FINAL TECHNICAL REPORT

ONR Award N00014-96-1-1296

Encoding of Complex Valued Composite Functions onto Spatial Light Modulators in Real-Time

University of Louisville and
Illinois Institute of Technology
7 December 2000

Report to: William J. Miceli, Office of Naval Research

Copy to: Director Naval Research Laboratory
ONR Administrative Grants Office, Chicago Regional Office
Defense Technical Information Center

Supplemental (electronic) copies to:

T. G. Baur, Meadowlark Optics
W. P. Bleha, Hughes JVC Technology Corp.
S. Block, Stanford University
G. W. Burr, IBM Almaden Research Center
D. M. Brown, MEMS Optical
D. R. Brown, MEMS Optical
T. H. Chao, NASA JPL
H. Cole, NASA MSFC
J. H. Comtois, AFRL
J. A. Davis, San Diego State University
F. M. Dickey, Sandia National Laboratory
M. Duelli, OCLI/JDS Uniphase
R. J. Feldmann, AFRL Wright
C. Filipietz, AFRL Phillips
D. H. Goldstein, AFRL Eglin
D. A. Gregory, University Alabama, Huntsville
D. G. Grier, University of Chicago
S. R. Harris, AFRL Wright
L. G. Hassebrook, University of Kentucky
J. Hines, NASA Ames Research Center
R. Hinkle, Army ECBC
D. A. Honey, DARPA MTO
L. J. Hornbeck, Texas Instruments
J. L. Horner, AFRL Hanscom
T. D. Hudson, U.S. Army Redstone Arsenal
W. R. Humbert, AFRL Eglin
R. D. Juday, NASA JSC
A. Keys, NASA MSFC
Y. Li, NEC Research Institute
R. Magnusson, University Texas, Arlington
J. N. Mait, ARL
P. F. McManamon, AFRL Wright
G. P. Nordin, University Alabama, Huntsville
D. W. Prather, University of Delaware
A. S. Rudolph, DARPA DSO
D. B. Searle, NAVAIR
H. Stark, Illinois Institute of Technology
J. E. Stockley, Boulder Nonlinear Systems
B. R. Stone, Mission Research Corporation
R. L. Sutherland, SAIC
K. Vaccaro, AFRL Hanscom
E. A. Watson, AFRL Wright
H. Willis, Army AMEDD
D. W. Wilson, NASA JPL
C. Woods, AFRL Hanscom
K. Wu, BMDO
M. H. Wu, Hamamatsu Corporation
F. T. S. Yu, Pennsylvania State University

REPORT DOCUMENTATION PAGE

Form Approved
OMB No. 0704-0188

1. AGENCY USE ONLY *(Leave Blank)*	2. REPORT DATE 7 DECEMBER 2000	3. REPORT TYPE AND DATES COVERED Final 1 September 1996 - 30 August 2000

4. TITLE AND SUBTITLE

Encoding of Complex Valued Composite Functions onto Spatial Light Modulators in Real-Time

5. FUNDING NUMBERS

N00014-96-1-1296

6. AUTHORS

Robert W. Cohn

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)

ElectroOptics Research Institute & Nanotechnology Center
Lutz Hall, Rm 442
University of Louisville
Louisville, Kentucky 40292

8. PERFORMING ORGANIZATION REPORT NUMBER

9. SPONSORING / MONITORING AGENCY NAME(S) AND ADDRESS(ES)

Office of Naval Research
Program Officer ONR William J. Miceli: 313
800 North Quincy Street
Arlington, Virginia 22217-5660

10. SPONSORING / MONITORING AGENCY REPORT NUMBER

11. SUPPLEMENTARY NOTES

12a. DISTRIBUTION / AVAILABILITY STATEMENT

Approved for public release; distribution unlimited

12b. DISTRIBUTION CODE

13. ABSTRACT *(Maximum 200 words)*

Spatial light modulators (SLM) are used as real-time programmable diffractive optics for potential applications to multi-object laser designation of moving objects. For real-time situations in which prior knowledge is unavailable, fast on-line design algorithms are required. The time requirements rule out many iterative design methods currently used for fixed pattern diffractive optics. Instead, encoding methods that map desired complex values onto the available modulation values provide the fastest realizations. As time permits, designs improved through limited iteration can be mapped to the SLM. Such a system, built around pseudorandom encoding (PRE) and its extensions was developed and experimentally demonstrated during the study.

The most significant extension to PRE is the development of a method of complex-valued encoding that can be accomplished with as few as three discrete modulation values, and which can be generalized to any modulator characteristic. The pairwise blending of PRE with three other encoding algorithms (minimum distance encoding-MDE, modified minimum distance encoding-mMDE, error diffusion-ED) provides one or two parameters that can be adjusted quickly to produce better performance than either algorithm produces individually. The new design methods were experimentally demonstrated together with demonstrations of real-time design and continuous scanning of multiple spots on arbitrary and independent trajectories.

14. SUBJECT TERMS

Spatial light modulators, diffractive optics, computer generated holography
Multi-spot beam steering, laser tweezer arrays, optical signal processing, statistical optics
Information optics, encoding, filter design, guided wave optics, multi-layer dielectric stacks

15. NUMBER OF PAGES

158

16. PRICE CODE

17. SECURITY CLASSIFICATION OF REPORT Unclassified	18. SECURITY CLASSIFICATION OF THIS PAGE Unclassified	19. SECURITY CLASSIFICATION OF ABSTRACT Unclassified	20. LIMITATION OF ABSTRACT UL

The ElectroOptics Research Institute
and Nanotechnology Center
Lutz Hall, Room 442

University of Louisville
Louisville, Kentucky 40292
(502) 852-7077
(502) 852-1577 (fax)
http://www.ee.louisville.edu/~eri

UNIVERSITYof LOUISVILLE

11 December 2000

William J. Miceli
Office of Naval Research
800 North Quincy Street
Arlington, VA 22217-5660

Dear Bill,

Per the contract requirements on N000014-96-1-1296 I am sending you a hard copy of the final report. I have already sent the report in electronic format by email to you and to several scientist who will be interested in the findings.

Hope your sabbatical went, or is still going well.

Best regards,

Bob

Robert W. Cohn, Ph.D.
Institute Director and Professor
Electrical and Computer Engineering

Attached: Final Report (3 copies)

CC: Director Naval Research Laboratory (1 copy of report)
ONR Administrative Grants Office, Chicago Regional Office (1 copy of report)
Defense Technical Information Center (2 copies of report)

ENCODING OF COMPLEX VALUED COMPOSITE FUNCTIONS ONTO SPATIAL LIGHT MODULATORS IN REAL-TIME

TABLE OF CONTENTS

The electronic copy does not contain this appendix.
However it does contain web links to pdf files of the papers.

ENCODING OF COMPLEX VALUED COMPOSITE FUNCTIONS ONTO SPATIAL LIGHT MODULATORS IN REAL-TIME

Robert W. Cohn (PI)
ElectroOptics Research Institute & Nanotechnology Center
University of Louisville

ABSTRACT

Spatial light modulators (SLM) are used as real-time programmable diffractive optics for potential applications to multi-object laser designation of moving objects. For real-time situations in which prior knowledge is unavailable, fast on-line design algorithms are required. The time requirements rule out many iterative design methods currently used for fixed pattern diffractive optics. Instead, encoding methods that map desired complex values onto the available modulation values provide the fastest realizations. As time permits, designs improved through limited iteration can be mapped to the SLM. Such a system, built around pseudorandom encoding (PRE) and its extensions was developed and experimentally demonstrated during the study.

The most significant extension to PRE is the development of a method of complex-valued encoding that can be accomplished with as few as three discrete modulation values, and which can be generalized to any modulator characteristic. The pairwise blending of PRE with three other encoding algorithms (minimum distance encoding-MDE, modified minimum distance encoding-mMDE, error diffusion-ED) provides one or two parameters that can be adjusted quickly to produce better performance than either algorithm produces individually. The new design methods were experimentally demonstrated together with demonstrations of real-time design and continuous scanning of multiple spots on arbitrary and independent trajectories.

The ONR award is:
R. W. Cohn, A. A. Vasiliev, and H. Stark (Illinois Institute of Technology), "Encoding of Complex Valued Composite Functions onto Spatial Light Modulators in Real-Time," Office of Naval Research, N00014-96-1-1296, $303,159. (1 September 1996 - 30 August 2000) *Note: Vasiliev relocated to Physical Optics prior to the start of the grant.*

STATEMENT OF THE PROBLEM

The properties of current and anticipated spatial light modulators (SLM's) limit the capabilities and performance of DoD systems in that they do not produce all complex values of modulation. SLM's can be amplitude-only, phase-only, or amplitude can even be a function of phase. However, practical devices providing independently controllable values of phase and amplitude, due to complexity and cost, are not expected for some time. It has been possible to design modulation patterns for these limited modulation range SLM's that do produce diffraction patterns similar to those possible from full complex SLM's, but hours of intensive iterative optimization can be required. Instead, we have studied methods of encoding that can be performed in real-time and partially optimized in near-real time, that make it possible to use SLM's in adaptive scenarios in which the lack of prior knowledge prohibits off-line design.

Potential applications include multi-target laser designators, active vision systems and vision based robotic navigation in which templates are projected and conformed onto objects to aid in recognition and tracking, optical switching and interconnects, and laser marking systems. Additionally the beam steering system has many potential biological and chemical sensing applications by attaching it to a light microscope. When the Fourier plane of the SLM is imaged onto the specimen the diffraction spots are at their tightest focus and can trap cell and particles. Multiple particles can be trapped, steered and brought together. If the particles are coated with specific materials it becomes possible to perform a series of tests, assays and even assemblies in a sequential fashion depending on the outcome of each test. Also the particles can be moved vertically by adding a defocus term (a spherical or quadratic phase profile to the spot modulation pattern.) The current limit to these and other possible real-time applications, has been the lack of a system that transparently couples the SLM and the diffractive design algorithms together in a real-time, multi-spot beam steering system. This report documents the progress made on this study towards making multi-spot systems a reality.

SUMMARY OF THE MAIN RESULTS

The results in this report are largely built on extensions to the pseudorandom encoding (PRE) method for phase-only SLM's that we originally developed on AFRL/DARPA grant F19628-92-K-0021 (DTIC report order number: ADA315727.) For this reason the section *Background* reviews encoding in general, PRE, and additional necessary concepts from diffractive optics design. Then the key results of the study are summarized together with web links to the corresponding journal papers and Quicktime movies. Hard copies of these papers are included in the *Appendix* of the non-electronic version. Additional results follow and are summarized similarly. Finally recommendations for future studies as suggested by this study are presented.

BACKGROUND

PRE is a statistically-based method of approximating fully complex modulation on SLM's that produce only a limited range of complex values (e.g. real valued, continuous phase-only over a full 2π, phase-only less than 2π, discrete valued phase, amplitude-phase coupled). Most encoding methods perform a single function operation per SLM pixel which leads to a numerically efficient computation that can even be performed in serial at SLM frame rates. PRE is notable in that it produces a good approximation to the desired function that one would obtain with a fully complex modulator, plus a faint white noise background. The noise level is directly related to the mean squared error between the desired complex modulation and the actual modulation produced by encoding. The noise background extends over the full usable diffraction plane (i.e. the non-redundant bandwidth-NRB which corresponds to the reciprocal of the pixel pitch.) There are no pronounced noise peaks or sidelobes as compared with prior methods.

The earliest computer generated holograms represented complex numbers by grouping pixels together, which results in regions of significant noise and which reduces the useful area of the diffraction plane. The method of minimum distance encoding (MDE) that was pioneered by Juday (which is often referred to as MEDOF-minimum euclidean distance optimal filter) minimizes the sum of the squared errors between the desired fully complex modulation and the actual modulation error. However, minimum distance errors in the modulation plane result in pronounced harmonic errors (large noise sidelobes) and reduced fidelity (e.g. increased deviations from a desired equal intensity spot array) in the diffraction plane. While the average noise background from MDE is lower than PRE, the intensity of the brightest noise sidelobe is usually greater than that found for PRE. These attributes of PRE often prove advantageous when overall fidelity across the entire NRB and speed of computation are key requirements.

The importance of complex valued encoding can be understood by considering how multi-spot beamsteering would be performed with a fully complex function and then by further considering how the performance is affected by using limited modulation range SLM's. Typically the SLM is used in a Fourier transform arrangement in which the SLM is illuminated by a collimated laser beam. The complex amplitude of the Fraunhofer diffraction pattern is the Fourier transform of the complex-valued modulation from the SLM. The Fourier transform is a linear operator. Therefore, when individual modulation plane functions are added together to form a composite modulation function, the resulting diffraction pattern is also found to be a linear combination of the Fourier transforms of the individual modulation plane functions. The fully complex SLM then provides a direct design procedure for synthesizing desired diffraction patterns. One needs only use known Fourier transform pairs and linear superposition to design any desired array of spots. This is a significant simplification over analytically performing the Fourier transform or numerically performing the fast Fourier transform (FFT).

However, if the SLM is not fully complex, mapping from the desired complex function to the actual SLM modulation values is a nonlinear operation, which is prone to producing undesirable interactions between the individual functions (specifically intermodulation distortion including sum and difference frequency generation.) In MDE, which uses hard decision boundaries to map from the fully complex to the SLM values and which is akin to hard limiting, the noise sidelobes are quite apparent at the sum and difference frequencies. In PRE the harmonics are not emphasized because the mapping is random rather than systematic. The SLM value is selected randomly in proportion to the closeness of the desired complex value to various possible SLM values. This soft decision method does not eliminate the noise harmonics, but tends to diffuse the sidelobes over the entire diffraction plane. On average, the expected value of the diffraction pattern is identically the desired diffraction pattern plus a white noise background that corresponds to the average error between the actual and the desired modulation.

An additional important issue in designing diffraction patterns is that of design freedoms. There is not one, but there are many possible modulation functions that produce identical intensity distributions in the diffraction plane. For example, consider an array of identical intensity spots in the diffraction plane. Each spot has an arbitrary value of phase. Thus, there are a near infinite number of possible complex amplitudes that produce the same intensity distribution. The values of these phase free parameters do profoundly affect the performance, even if the modulator is fully complex. Specifically, the diffraction efficiency can vary from $1/N$ where N is the number of spots for all phases identical, to 70% or higher for an optimal selection of the phases. In the full complex SLM the only major concern is the loss of energy, which is absorbed by the SLM.

However, for a limited modulation range SLM, error between the desired function and the encoded function also produces a noise background and nonuniformity errors in the resulting diffraction pattern. (And for a phase-only SLM none of the error signal is absorbed and all the error signal reaches the diffraction pattern.) For the limited-range SLM, optimization in terms of the free parameters can also be carried out. In these designs diffraction efficiencies of 90% to nearly 100% have been produced (for phase only modulators.) The designs usually include an allowance for a few percent intensity nonuniformity of the spot array. The high efficiency helps to reduce the background noise.

While the performance of these optimized designs is exceptional (as routinely demonstrated by the manufacturers of fixed pattern diffractive optics), the computational load can be impractical for a multi-spot beam steering system that is designing and projecting a sequence of diffraction patterns in real-time. These computational issues led to the hierarchical design approach in [1] (specifically, Fig. 1) in which successively more numerically intensive design methods are employed depending on the available computational time, which in the real-world is situation dependent and varies from moment to moment. Review articles [19,20] cover the background in this section in greater depth.

KEY RESULTS FROM THE STUDY:

Result 1: PRE for Ternary, discrete and arbitrary SLM's. The method of PRE which had previously been applied only to SLM's that are phase-only or phase-only plus an added zero amplitude states, was extended to a completely general form in [10]. The method requires that the SLM produce at least three complex values. The probability of selecting each of the three values is determined by solving a third order linear equation. The range encoded by PRE is contained within the triangular region formed by connecting the three values. For discrete SLM's having more than three values, multiple triangular regions can be defined in order to enlarge the encodable range or to more finely subdivide the complex plane. The imposition of smaller triangular regions leads to reduced encoding errors and hence higher fidelity diffraction patterns.

The practical benefit of the three level encoding method is that one could develop prototype optical processors with three value SLM's. The simplified requirements on addressing would greatly accelerate development and reduce cost of the SLM. The value of developing an SLM with a greater number of modulation values could then be estimated based on the resulting reduction in encoding error. Application of the new PRE method to continuous-valued SLM's is then based on a practical assessment of the number of modulation values that need to be addressed.

Result 2: Blended PRE algorithms. The concept of combining multiple encoding algorithms together was explored and extended. While PRE has a limited encoding range, MDE (and also error diffusion – ED, which is discussed in *Result 8*) encode any value on the complex plane. The so-called *blended* algorithms use MDE (or ED) to encode those values that cannot be encoded by PRE.

The values encoded by either PRE or MDE can be changed by simply scaling the desired complex values in amplitude and/or phase. This provides two free parameters that can be used to fine tune the fidelity of the diffraction pattern. This tuning usually results in a blended algorithm that produces a diffraction pattern of higher fidelity than either algorithm produces individually. For background noise, it is found that increased encoding by MDE reduces the random noise of PRE but increases the harmonic noise of MDE. There is usually a clearly evident degree of blending for which the combined noise level from the two types of encoding is minimized. Similarly there is a blending that produces a minimum uniformity fluctuation from the desired diffraction pattern due to reduced contributions of the random fluctuations from PRE and intermodulation distortion from MDE. Typical plots of the performance changes as a function of the degree of blending are shown in Fig. 8 of [5]. At present there is no numerically efficient procedure or approximate method for selecting the blending free parameters optimally. Multiple simulations, each involving a Fourier transform, is currently required. However, it would be fairly easy to perform the adjustment using a video camera to feed back the far-field intensity patterns.

The most interesting and novel result of this paper was the proposal and development of a modified MDE-PRE (MD-PRE) blended algorithm, that we refer to as modified MD-PRE (mMD-PRE). The MDE prescription is to map the desired complex value to the closest value produced by the SLM. The mMDE prescription is to map the desired value to the closest value produced by PRE and then use PRE to encode the mapped value. (See Fig. 1 in [5].) These two distinct possibilities were not originally apparent to us until we began considering discrete value SLM's. The reason the distinction was not apparent is that in our original studies the minimum distance mapping happens to coincide with the modified minimum distance mapping for continuous phase-only SLM's. Simulated diffraction patterns (Fig. 9 from [5]) for a three phase, phase-only SLM provides the clearest demonstration and comparison of MDE, PRE, MD-PRE and mMD-PRE. The mMD-PRE demonstrates the highest fidelity of the four encoding methods, and is clearly seen to reduce harmonics over the MD-PRE method.

A special case of mMD-PRE for real-valued ternary SLM's is considered in [8]. Similar observations on the performance improvements of mMD-PRE and MD-PRE over MDE and PRE individually are made. The real-valued modulation necessarily produces diffraction patterns that are symmetric around the optical axis. Typically the symmetry implies a reduction in the useable non-redundant bandwidth. However, in some applications (e.g. for specific optical network topologies) symmetric diffraction patterns may be desirable.

Result 3: Optimized selection of the fully complex function for encoding. As reviewed in *Background* on the phase free parameters, there are many fully complex functions that produce identical intensity distributions. For PRE onto a phase-only SLM, it is desirable to select the fully-complex function that has the highest diffraction efficiency (η). Since η can be directly calculated from the complex valued modulation (specifically, η is the average intensity modulation across the SLM aperture) optimization of η can be performed without iterative use of the FFT, which is likely to provide numerical advantages over other design methods that iteratively transform between the modulation plane and the diffraction plane. Three methods of optimization (Monte Carlo, genetic algorithm, and gradient descent) were applied to the problem of identifying a fully complex function that has the highest diffraction efficiency. The results of the study for a 10 spot pattern are summarized in Table 2 of [4]. The gradient method achieved the highest efficiency (η = 69.3%) with the least computational effort, while the genetic algorithm (η = 64.7%) achieved the most uniform diffraction pattern.

To put these results in perspective, Table 2 also reports the performance for setting the amplitude of each modulation value to unity. This corresponds to the classical kinoform and also to MDE for a phase-only SLM. This phase-only transformation is known to produce the highest diffraction efficiency. The efficiency of the kinoform for the modulation designed by the genetic algorithm actually achieves the theoretical maximum possible efficiency (η = 98%) over the complete space of the phase free parameters (as reported in Krackhardt *et al.*, *Appl. Opt.* **31**, 27-37, 1992.) The efficiency is within 1.5% of the theoretical limit for the kinoform design based on the gradient algorithm. Even for the 1000 iteration Monte Carlo optimization of the fully complex function (η = 47.7%) the kinoform efficiency is within 3.8% of the maximum possible efficiency. The wide variation in η for the fully complex function reduces to a much smaller variation in η for kinoforms.

This result has some implications for selection of fully complex functions for blended algorithms, such as MD-PRE, in a real-time system. A question raised is whether more computational effort should be devoted to optimizing the efficiency of the fully complex function or to searching for the optimal degree of blending. This problem could be further investigated by developing a table similar to Table 2 for the case of MD-PRE. Table 1 in [4] reports the optimum values of phase for each optimization method, and would permit this question to be investigated directly. (So far, in most of our journal papers on blended encoding [2,5,8], we only have used the free parameters of phase derived from Krackhardt *et al.* that produce near the theoretical maximum efficiency for the kinoform.) A study using various fully complex input efficiencies would be an important step towards developing a numerically efficient algorithm for simultaneously optimizing the phase and the blending free parameters.

Result 4: Experimental demonstrations of beamsteering with phase-only SLM's. Experimental demonstrations were performed with a Hughes liquid crystal light valve (LCLV) and a Boulder Nonlinear Systems (BNS), 128x128 pixel electrically-addressed SLM filled with parallel-aligned nematic liquid crystal. Both SLM's produce essentially continuous phase modulation over a 2π range. The BNS SLM behaves more like an array of independently addressed phase-only elements than does the LCLV (which suffers from its limited resolution. This topic is discussed further in *Result 9*.) For the BNS SLM, the most significant difference between the device an ideal SLM is the presence of a bright on-axis spot which is due to Fresnel reflections from the cover glass and which is further enhanced by the low reflectance of the modulating layer. Experimental results are reported in several of the papers, and in all cases the experimental results compare quite reasonably well with the simulated results. In Tables 1,2 and Fig. 13 of [5] theory and experiment are compared for blended encoding for continuous, three-phase and four-phase phase-only SLM's. The greatest discrepancy seems to be that the measured level of non-uniformity (especially in Fig. 13) seems to be offset to a somewhat higher level than for theory. Nonetheless, the same trends are found for nonuniformity as a function of the degree of blending. Also this SLM was applied to simulate the effect of blurring observed in the LCLV (through digital predistortion of the addressing signal) in [3] (as described in *Result 9*.)

The SLM was used extensively to demonstrate real-time multi-spot beam steering in [1]. The general concepts of arbitrary beam steering are illustrated by the Quicktime movies associated with Fig. 2a and Fig. 3a in [1]. MD-PRE is compared with MDE in Fig. 3a and Fig. 3b. The distinct, harmonically-related sidelobes of MDE should be more easy to falsely identify as a target than the more uniform, matte-finish noise background of MD-PRE. Fig. 4 shows continuous translation of a fixed pattern (a 7x7 spot array).

Fig. 5a shows that continuous, gap-free scanning, similar to a galvanometric scanner is possible with several spots simultaneously. Fig. 5b illustrates the problems of using an FFT-based design algorithm–which is that the spots are located at discrete locations, as compared to the continuous locations possible using a Fourier transform table lookup approach as was used for Fig. 5a. Fig. 8b in [1] illustrates that nonuniformity can be reduced (when time permits) by averaging together sequential realizations (Fig. 8a) of the same diffraction pattern. Fig. 9 in [1] demonstrates that the SLM can be used to provide broad-area laser illumination of objects (in this case a coin.) A broadly spread pattern is produced by randomly selecting phases for the SLM pixels from a uniform distribution over a 2π range. The resulting speckle pattern is reduced to acceptable levels in Fig. 9b by averaging together successive realizations. Fig. 7a illustrates independent shaping of several beams in parallel by using aperture subdivision together with PRE to apodize the sub-apertures of the SLM subdivision. Fig. 7b generalizes the previous example by adding a second layer of modulation functions to the previous modulation function, resulting in three more spots. This example suggests a controller that assigns an area on the SLM each time a new spot is needed. The controller assigns space based on trying to optimize overall performance.

Fig. 6a demonstrates a numerically efficient and diffraction efficient way to scan many spots in parallel. In this movie one cluster of 16 spots corresponds to 16 interleaved functions (each a linear phase ramp). The undersampling of each function (each function is sampled once in a 4x4 array of SLM pixels) results in a 4x4 replication of the 16-spot, spot array. The entire pattern has a theoretical diffraction efficiency of 100%. The replications can be turned off by random (rather than periodic) multiplexing of the 16 switching functions as shown in Fig. 6b. The random multiplexing method was proposed by (Davis and Cottrell, *Opt. Lett.* 19, 496-498, 1994). These experiments demonstrate and suggest an extremely wide range of functions for beam steering and laser illumination that are possible using a SLM together with an intelligent on-line design system.

Result 5: System specification of an on-line diffractive design system for multi-spot beemsteering. In order to provide the fastest throughput of multi-spot designs we proposed the hierarchical design method described in Fig. 1a in [1]. Successively more numerically intensive design methods are employed depending on the available computational time. The first level design method is to specify and compose a fully complex function and then encode it by a selected method e.g. MD-PRE. A few free parameters (e.g. scaling of the fully complex function) can be used to improve on the initial encoding in the second level of the design. The third level is to use the free parameters associated with the spot phases to improve diffraction efficiency. Along with the free parameters of phase, spot intensities can be adjusted iteratively over a small range to compensate for the resulting deviations from the desired intensity levels. For this system, the third level of the hierarchy is only executed as time permits. If even more time is available, then the direct optimization methods for fixed pattern diffractive optics (e.g. the iterative Fourier transform in Fig. 1b) can even be employed. As shown by the results in Sec. 4, good patterns often can be produced by using only the second, or even the first level.

In summary, the key result of [1] is the design of an on-line modulation pattern design system that together with the phase-only SLM demonstrates adaptive beam steering of arrays of individually and continuously steerable laser spots.

ADDITIONAL RESULTS FROM THE STUDY:

Result 6: PRE for amplitude-phase coupled SLM's. Even SLM's that are claimed to be phase-only, usually exhibit amplitude variations as a function of phase. Furthermore the use of polarizers together with twisted or parallel-aligned liquid crystal SLM's can result in spiral and off-center circular modulation curves on the complex plane. The development of PRE methods for coupled SLM's was first addressed in [11]. The paper presents several algorithmic approaches and identifies various limitations related to realizability of desired values, numerical efficiency, and encoding errors. It is shown that various PRE encoding methods can realize the same complex value, but with differing levels of encoding error. Also it is shown that many encoding algorithms reduce the encoding range to less than the maximum possible range for PRE. The range appears to be maximized using a method of binary selection between two possible values on the modulation curve. (Also see *Result 7.*)

Result 7: Systematic evaluation of the full encoding range that can be encoded by PRE. This paper [9] specifically focuses on and builds on the recognition from [11] that the binary encoding method provides a way to determine the encoding range of PRE. Binary encoding is shown to encode any value on the line segment that connects two modulation values by simply adjusting the probability of randomly selecting each value. Line segments can be iteratively drawn between all pairs of modulation values to map out the fully PRE encoding range. This evaluation produces a convex set of values. Within the convex region one can identify the fully complex encoding range as the largest range that can be enclosed by a circle around the origin of the complex plane. Finally the possibility of ternary PRE is mentioned. Further consideration of this possibility led to ternary encoding as described in *Result 1.*

Result 8: Blended encoding of PRE with error diffusion. Error diffusion (ED) adds a nonlinear filtering operation to the MDE algorithm. The method maps a desired complex value to the closest available modulation value of the SLM. The error between the desired and actual value is then weighted (by a linear filtering coefficient) and added onto the next complex value to be encoded. The error adjusted complex value is again encoded by MDE and the next value of error is calculated. The performance of ED has been impressive. However the method, due to the inherent filtering in the algorithm, does tend to produced pronounced noise outside the window of the desired diffraction pattern (but still within the non-redundant bandwidth of the diffraction pattern.) Blending of ED with PRE (ED-PRE) was demonstrated in [2], specifically for continuous phase-only modulation characteristics. As with ED, there is a filtered value that is encoded. Values outside the unit circle are encoded by ED. Values inside the unit circle are encoded by PRE. In addition to amplitude scaling of the entire complex function (as in MD-PRE) a new free parameter is included that scales the amount of the error that is diffused forward when PRE is selected. This new factor accounts for

for the fact that PRE already distributes error into a white background of noise, thus propagating forward all the error would actually introduce more encoding error than is necessary. Searching over the two free parameters led to diffraction patterns that had higher fidelity than either method produces individually. ED-PRE tended to replace harmonic noise peaks with speckle patterns that had lower levels of peak noise. The noise pattern was not white, but instead produced a reduced noise region around the desired diffraction pattern. Overall, the fidelity of ED-PRE can sometimes be somewhat better than for MD-PRE. The small performance improvement may not justify the search over the extra parameter. However, [2] does show that there are ways that ED (which is widely used in diffractive optic design and printer halftoning) might be improved through blending.

Result 9: Nonlinear distortion of diffraction patterns due to blurring of the phase modulation. The effects from the limited resolution of current liquid crystal light valves (LCLV's) appears to be much more severe when these SLM's are used in a phase-only mode for producing Fraunhofer diffraction patterns, than when they are used as intensity modulators for image display and projection [3]. A typical resolution specification for commercial LCLV's is 40 lp/mm for low intensity write signals and 4 lp/mm for high level write signals. When we measured the resolution of the phase modulation directly (rather than the resolution on an intensity image) we found that this resolution is unchanged. The problem is that increasing depth of modulation, leads to increasing nonlinear distortion, and a large depth of modulation (up to 2π) is required for diffraction-efficient beam steering. The nonlinearity becomes evident when one convolves the point spread function (PSF) of the phase modulation with a desired signal. This blurring PSF (which we measured to have a FWHM of ~60 μm FWHM for a Hughes LCLV) when convolved with the ideal write signal and Fourier transformed reproduced all the harmonic features observed in the measured diffraction pattern. Further simulations of the performance of a 128x128 pixel PRE design showed that the width of PSF needs to be 5% of the pixel spacing to maintain levels of uniformity and harmonics reasonably close to the levels anticipated without phase blurring. While predistortion of the write signal can be envisioned to compensate for the blurring, we found that predistortion requires that the SLM have an increased phase range (often many times greater than 2π) to produce the correction.

The best solution to the problem appears to be minimizing the phase blurring in LCLV's in the first place. The blurring is probably due to electrical field fringing between the parallel plate electrodes of the SLM. Blurring was not observed in the BNS SLM, for which each pixel is electrically shielded from each other. Therefore, for LCLV's a reasonably practical solution would be to fabricate a pixelated device in which each pixel is electrically isolated from all others. We anticipate that pixels as small or smaller than 3 μ could be fabricated. The fabrication requirements would be much less involved than making an electrically-addressed SLM. Imaging and reduction imaging systems for imaging the write signal with resolutions of 3 μm would not be technically challenging to develop. The pixelated LCLV would have the advantages of being able to handle very intense laser illumination, much larger pixel counts and much wider diffraction angles than current electrically-addressed SLM's. The manufacturing costs and time could be much less than for electrically-addressed SLM's. These features would also make the LCLV attractive for optical interconnects and optical backplanes.

Result 10: Reduction of coherent interference on laser diffraction patterns. We developed a measurement procedure that uses a frequency-swept laser diode to reduce unwanted interference fringes noise in diffraction pattern measurements due to multiple reflections in [12]. While CCD cameras can be purchased without cover glass over the imager, often cooled CCD cameras cannot. Reflections between the imager and the cover glass produce interference fringes, that make it difficult to accurately measure the diffraction pattern. One way to reduce this effect is to use a temporally incoherent source of light. However, with the recent availability of tunable laser diodes one can instead sweep the laser diode and time integrate the diffraction pattern, which averages out the interference fringes. We measured the pattern from a glass diffractive optic by this method. The resulting 8x8 spot array had a theoretical nonuniformity of 7% rms and the measured uniformity was 7.9% with the laser swept over a 0.25 nm range. Without sweeping the uniformity was measured as 12.1%. Current tunable laser diode systems can sweep adequately fast to permit cancellation of coherent interference at video frame rate.

Result 11: Applications of encoding to other device technologies. PRE and its derivatives can be applied to device technologies other than SLM's. PRE could be used in a variety of phased array systems. For instance, acoustic direction finding arrays consist of many individual acoustic sensors whose directionality is steered by phasing of the sensors. Using PRE with a three phase, phase shifter would effectively permit continuous, multi-directional scanning of directivity. The ternary phase shifter would likely have cost and size advantages over a continuous shifter.

A second application would be the rapid design and fabrication of fixed-pattern diffractive optics in [13]. Fabrication speed results from using laser speckle to photolithographically expose custom texture patterns on each pixel of a diffractive optic. This type of lithography can be much faster than direct-write with a focused electron or laser beam. The roughness pattern across a pixel can be viewed as a repeated random trials or spatially multiplexed realizations of the particular PRE design. Repeated trails provide more averaging than a single trial per pixel (as is in PRE on SLM's), which results in improved performance. While the error signal is identical, it is diffracted over a greater spatial extent than for a single random trial, resulting in a lower average noise level. Most of the research reported in this paper was completed prior to the current grant.

A third application conceived on this study was the idea of using PRE in a Bragg grating on top of a waveguide (or similarly, in a multi-thin film dielectric stack.) By modulating the position of Bragg grating stripes from their normal periodic locations, it becomes possible to design multi-passband reflection filters. Fabrication approaches that could lead to rapid prototyping of custom filters is also described in [7]. Fabrication development of the device is the subject of current grants.

RECOMMENDATIONS FOR FUTURE STUDIES:

Recommendation 1. Per *Result 5*, develop a portable prototype multi-spot beamsteering system for demonstration to interested military, biomedical and commercial parties. Some work has been begun on a Phase I STTR grant with Boulder Nonlinear Systems, Inc.

Recommendation 2. Per *Result 5*, demonstrate an adaptive multi-object tracking and laser designation system in which video tracking software provides a set of coordinates to the on-line modulation design system and the resulting modulation positions laser spots on the moving objects. A skeleton system has been begun on the Phase I STTR grant. L. G. Hassebrook of U. of Kentucky has been leading the multi-object tracking software development activity.

Recommendation 3. Continue experimental demonstrations with new or improved SLM's as they become available including the development of encoding algorithms specific to the modulation characteristics of the device.

Recommendation 4. Per *Result 9*, develop an optically-addressed SLM's or light valve that through use of a non-addressed, pixelated SLM, is immune to phase-blurring down a resolution set by the pixel size of the SLM and the resolution of the image projected on the write side of the light valve.

Recommendation 5. Fabricate and experimentally demonstrate waveguide grating devices designed using PRE per *Result 11*. This work is ongoing in current and pending grants.

LIST OF PARTICIPANTS

Participants University of Louisville (UofL)
Robert W. Cohn, PI, Advanced encoding method & theory, Architect of real-time design system
Markus Duelli, Ph.D., Postdoctoral Scientist for most of the study, Lead experimentalist and designs
Sergei F. Lyuksyutov, Ph.D., Postdoc. Simulations of the PRE-based waveguide grating filter
Li Ge, GRA, Received MS on University Fellowship, Major experimental and design contributions
David L. Hill, GRA, Received MS under ASSERT funding, Experimental contributions
Christy L. Lawson, Received MS under ASSERT funding, Contributed to designs & simulations
Manivannan S. Vangalur, Received MS (self-supported), Contributed to designs and simulations
Matthew Reece, High School Intern, Studies on mMD-PRE. (Awarded 6th in Intel Talent search)

Participants Illinois Institute of Technology (IIT)
Henry Stark, Subcontract PI, Advanced methods and theory of designing functions for encoding
Yongyi Yang, Asst. Prof., Developed theory and simulations of designing functions for encoding
Damla Gurkan, GRA, Completed Ph.D. on this study, Computer evaluations of design methods

REFEREED PUBLICATIONS (with web links)

[1] L. Ge, M. Duelli, and R. W. Cohn, "Enumeration of illumination and scanning modes from real-time spatial light modulators," *Optics Express* 7(12), 403-416 (4 December 2000) http://www.opticsexpress.org/oearchive/pdf/26795.pdf

[2] L. Ge, M. Duelli, and R. W. Cohn, "Improved fidelity error diffusion through blending with pseudorandom encoding," *Journal of the Optical Society of America A* 17(9), 1606-1616. (September 2000) http://www.ee.uofl.edu/~eri/papers_pres/josaed.pdf

[3] M. Duelli, L. Ge, and R. W. Cohn, "Nonlinear effects of phase blurring on Fourier transform holograms," *Journal of the Optical Society of America A* 17(9), 1594-1605. (September 2000) http://www.ee.uofl.edu/~eri/papers_pres/josablur.pdf

[4] Y. Yang, H. Stark, D. Gurken, C. L. Lawson, and R. W. Cohn, "High-diffraction-efficiency pseudorandom encoding," *Journal of the Optical Society of America A*, **17**(2), 285-293. (February 2000) http://www.ee.uofl.edu/~eri/papers_pres/stark_old.pdf

[5] M. Duelli, M. Reece, and R. W. Cohn, "Modified minimum-distance criterion for blended random and nonrandom encoding," *Journal of the Optical Society of America A*, **16**(10), 2425-2438. (October 1999) http://www.ee.uofl.edu/~eri/papers_pres/reece99.pdf

[6] J. Gotpagar, S. F. Lyuksyutov, R. W. Cohn, E. A. Grulke, and D. Bhattacharyya, "Reductive dehalogenation of trichloroethylene with zero-valent iron: Surface profiling microscopy and rate enhancement studies," *Langmuir* **15**, 8412-8420. (web version released 18 September 1999) http://www.ee.uofl.edu/~eri/papers_pres/langmuir.pdf

[7] R. W. Cohn, S. F. Lyuksyutov, K. M. Walsh, and M. M. Crain, "Nanolithography considerations for multi-passband grating filters," *Optical Review*, **6**(4) 345-354. (July/August 1999) http://www.ee.uofl.edu/~eri/pages/wdmpapver2.pdf (The link is to the manuscript only. Hard copies of the journal reprint are available on request.)

[8] M. Duelli, and R. W. Cohn, "Pseudorandom encoding for real-valued ternary spatial light modulators," *Applied Optics*, **38**(17) 3804-3809. (10 June 1999) http://www.ee.uofl.edu/~eri/papers_pres/ao99.pdf

[9] R. W. Cohn, "Analyzing the Encoding range of amplitude-phase coupled spatial light modulators," *Optical Engineering*, **38**(2), 361-367. (February 1999) http://www.ee.uofl.edu/~eri/papers_pres/opteng.pdf

[10] R. W. Cohn and Markus Duelli, "Ternary pseudorandom encoding of Fourier transform holograms," *Journal of the Optical Society of America A*, **16**(1), 71-84 and "errata," **16**(5) 1089-1090. (January 1999) http://www.ee.uofl.edu/~eri/papers_pres/c&d1999.pdf

[11] R. W. Cohn, "Pseudorandom encoding of complex-valued functions onto amplitude-coupled phase modulators," *Journal of the Optical Society of America, A*, **15**(4), 868-883. (April 1998) http://www.ee.uofl.edu/~eri/papers_pres/c1998.pdf

[12] M. Duelli, D. L. Hill and R. W. Cohn, "Frequency swept measurements of coherent diffraction patterns," *Applied Optics*, **37**(34), 8131-8133. (1 December 1998) Originally appeared in *Engineering & Laboratory Notes*, 3-5. Supplement to *Optics and Photonics News* **9**(2). (February 1998) http://www.ee.uofl.edu/~eri/papers_pres/frequency_swept.pdf

[13] R. W. Cohn, A. A. Vasiliev, W. Liu and D. L. Hill, "Fully complex diffractive optics by means of patterned diffuser arrays: Encoding concept and implications for fabrication," *Journal of the Optical Society of America, A*, **14**(5), 1110-1123. (May 1997) http://www.ee.uofl.edu/~eri/papers_pres/1110_1.pdf

PATENTS AND INVENTIONS

[14] R. W. Cohn, "Multi-spot beamsteering system and applications thereof," (4 December 2000, provisional patent applied for)

[15] R. W. Cohn and Li Ge, "Method of producing complex valued optical modulation from limited range light modulating arrays by blending error diffusion with pseudorandom encoding," (20 December 1999, Invention disclosure to NASA GSFC for possible U.S. patent filing.)

[16] R. W. Cohn, "Nanolithography for multi-passband grating filters," (U.S. patent pending, 09/575252, 22 May 2000)

[17] R. W. Cohn, "Modified minimum distance criterion for blended random and nonrandom encoding," (U.S. patent pending, 09/575270, 22 May 2000)

[18] R. W. Cohn, "Method of producing continuous optical modulation with discrete level or other non-continuous light modulating arrays," (U. S. Patent Pending, Ser. No. 09/301788, 29 April 1999)

BOOK CHAPTER

[19] R. W. Cohn and L. G. Hassebrook, "Representations of Fully Complex Functions on Real-Time Spatial Light Modulators," Ch. 15, pp. 396-432 in *Optical Information Processing*, F. T. S. Yu and S. Jutamulia, eds. Cambridge U. Press. (1998)

FULL-LENGTH CONFERENCE PROCEEDINGS

[20] R. W. Cohn, "Fundamental properties of spatial light modulators in Fourier transform applications," invited talk to *Symposium on Photonic Devices and Algorithms for Computing II*, K. M. Iftekharuddin and A. S. Awwal, eds., *Proc. SPIE.* (2 August 2000, San Diego, CA)

[21] D. L. Hill, M. Duelli and R. W. Cohn, "Improved spot-array generation using diffuser pixels to represent complex valued modulations," *Diffractive/Holographic Technologies and Spatial Light Modulators*, R. L. Sutherland, ed., *Proc. SPIE 3633*, 260-268. (29 January 1999, San Jose, CA)

[22] R. Cohn and W. Liu, "Encoding of fully-complex values onto amplitude-phase coupled spatial light modulators," *Proc. OII '98, Optics for Information Infrastructure.* G. G. Mu, ed., *J. Optoelectronics, Laser (JOEL)*, Vol. 9, supp., pp. 325-327. (6 August 1998, Tianjin, China)

[23] R. W. Cohn, "Analyzing the encoding range of amplitude-phase coupled spatial light modulators," *Spatial Light Modulators,* R. L. Sutherland, ed. *Proc. SPIE 3297,* 122-128. (29 January 1998, San Jose, CA)

[24] R. W. Cohn, "Real-time multispot beam steering with electrically controlled spatial light modulators," *Optical Scanning Systems: Design and Applications,* L. Beiser and S. F. Sagan, eds. *Proc. SPIE 3131,*145-155 (30 July 1997, San Diego, CA)

TALKS AT NATIONAL CONFERENCES

[25] R. W. Cohn, M. Duelli, and M. Reece, "A modified minimum distance criterion for blended random and nonrandom encoding," *Optical Society of America Annual Meeting*, Santa Clara, CA, paper WLL60. (29 September 1999)

[26] R. W. Cohn, S. F. Lyuksyutov, M. M. Crain, S. Sharma, P. Koduri, K. M. Walsh, and M. K. Sunkara, "Nanolithography considerations for multi-passband grating filters," *Symposium on Trends in Guided Wave Devices. Optical Society of America Annual Meeting*, Santa Clara, CA, paper WJ6. (29 September 1999)

[27] R. W. Cohn and M. Duelli, "Fully complex encoding of quantized spatial light modulators," *Symposium on Pattern Recognition: 5 - Hardware. Optical Society of America Annual Meeting*, Baltimore, MD, paper ThSS2. (8 October 1998)

[28] M. Duelli and R. W. Cohn, "Nonlinear effect of resolution loss on phase-only spatial light modulators," *Symposium on Pattern Recognition: 5 - Hardware. Optical Society of America Annual Meeting*, Baltimore, MD, paper ThSS4. (8 October 1998)

[29] R. W. Cohn, S. F. Lyuksyutov, M. M. Crain and K. W. Walsh, "Nanofabricated gratings for wavelength division multiplexing," *Gordon Research Conference on Chemistry and Physics of Nanostructure Fabrication*, Tilton, NH. (22 June 1998)

[30] R. W. Cohn, "Rationale for pseudorandom encoding of real-time spatial light modulators," Invited Talk. *Gordon Research Conference on Optical Signal Processing and Holography*, Meriden, NH. (3 July 1997)

SEMINARS AND TALKS AT REGIONAL MEETINGS

[31] R. W. Cohn, "Maskless three dimensional device lithographies," to Huntsville ElectroOptics Society. (15 July 1999)

[32] R. W. Cohn, "Statistically-based encoding of full complex values onto partially complex spatial light modulators," Optical Research Associates, Pasadena CA. (28 July 1997)

[33] R. W. Cohn, "Statistically-based encoding of full complex values onto partially complex spatial light modulators and applications," Army Research Laboratory, Adelphi, MD. (15 July 1997)

[34] R. W. Cohn, "Real-time beam steering functions with electrically-controlled spatial light modulators," Wright Laboratory, Avionics Lab, Dayton, OH. (12 December 1996)

APPENDIX

Thirteen journal papers funded or partially funded by this study are reprinted here in the hardcopy version. Web links are provide in the body of the text for those receiving the electronic version.

Enumeration of illumination and scanning modes from real-time spatial light modulators

Li Ge, Markus Duelli and Robert W. Cohn
ElectroOptics Research Institute, University of Louisville
Louisville, KY 40292 USA
rwcohn@louisville.edu

Abstract: Using a phase-only spatial light modulator (SLM) in a Fourier transform setup together with fast diffractive optics design algorithms provides a way to automatically generate complex and rapidly changing laser illumination patterns in the far-field. We propose a hierarchical software structure for the adaptive, on-line design of far-field illumination patterns. Using the on-line design system together with camera feedback of the illuminated scene would make it possible to detect and actively laser designate multiple objects in parallel. Possibilities for multispot, arbitrary trajectory scanning and also broad-area speckle-reduced illumination are demonstrated with experimentally measured diffraction pattern sequences from a 120 x 128 pixel phase-only SLM.

OCIS codes: (070.2580) Fourier optics; (230.6120) spatial light modulators; (090.1760) computer holography; (090.1970) diffractive optics; (030.6600) statistical optics

References and links

1. U. Krackhardt, J.N. Mait, and N. Streibl, "Upper bound on the diffraction efficiency of phase-only fanout elements," Appl. Opt. **31**, 27-37 (1992).
2. R. W. Cohn and L. G. Hassebrook, "Representations of fully complex functions on real-time spatial light modulators," Ch. 15, pp. 396-432 in *Optical information processing*, F. T. S. Yu and S. Jutamulia, eds. (Cambridge U. Press., Cambridge, UK, 1998).
3. Galvanometer scanners and accessories, http://www.gsilumonics.com/c03oem_gal_frame/galvoframe.html
4. Cambridge Technology Products, http://www.camtech.com/prods4a.htm
5. NEOS online catalog, http://www.neostech.com/neos/catalog
6. Isomet Corporation deflectors, http://www.isomet.com/deflectors.html
7. Boulder Nonlinear Systems, Inc., http://www.bnonlinear.com
8. T. H. Lin, "Implementation and characterization of a flexure-beam micromechanical spatial light modulator," Opt. Eng. **33**, 3643-3648 (1994).
9. M. Duelli, M. Reece, and R. W. Cohn, "Modified minimum distance criterion for blended random and nonrandom encoding," J. Opt. Soc. Am. A **16**, 2425-2438 (1999).
10. M. Duelli, R. W. Cohn, "Pseudorandom encoding for real-valued ternary spatial light modulators," Appl. Opt. **38**, 3804-3809 (1999).
11. L. Ge, M. Duelli, and R. W. Cohn, "Improved-fidelity error diffusion through blending with pseudorandom encoding," J. Opt. Soc. Am. A **17**, 1606-1616(2000).
12. R. W. Cohn and M. Liang, "Approximating fully complex spatial modulation with pseudo-random phase-only modulation," Appl. Opt. **33**, 4406-4415 (1994).
13. R. W. Cohn and M. Duelli, "Ternary pseudorandom encoding of Fourier transform holograms," J. Opt. Soc. Am. A **16**, 71-84 and "errata," 1089-1090 (1999).
14. R. D. Juday, "Optimal realizable filters and the minimum Euclidean distance principle," Appl. Opt. **32**, 5100-5111 (1993).
15. Y. Yang, H. Stark, D. Gurken, C. L. Lawson, and R. W. Cohn, "High-diffraction-efficiency pseudorandom encoding," J. Opt. Soc. Am. A **17**, 285-293 (2000).
16. J. N. Mait, "Understanding diffractive optic design in the scalar domain," J. Opt. Soc. Am. A **12**, 2145-2158 (1995).
17. D. Jared and D. Ennis, "Inclusion of filter modulation in synthetic discriminant function construction," Appl. Opt. **28**, 232-239 (1989).
18. N.C. Gallagher and B. Liu, "Method for computing kinoforms that reduces image reconstruction error," *Appl. Opt.* **12**, 2328-2335 (1973).
19. R. W. Cohn, "Adaptive real-time architectures for phase-only correlation," Appl. Opt. **32**, 718-725 (1993).
20. M. S. Vangular, *Optimization methods for diffraction gratings and composite matched filters*, M.S. Thesis,

Received October 17, 2000; Revised November 25, 2000

University of Louisville (1998).
21. J. A. Davis and D. M. Cottrell, "Random mask encoding of multiplexed phase-only and binary phase-only filters," Opt. Lett. **19**, 496-498 (1994).

1. Introduction

Spatial light modulators (SLM) can be used both as image displays and as programmable real-time diffractive optical elements (DOE). The second application is the area addressed by this paper. Programmable diffraction has a decided advantage over image display when the image consists of only a few bright pixels. Consider that when an SLM of N pixels is imaged, only ~1/N of the energy in the SLM aperture is found in one pixel of the image. However, when laser light is diffracted, as much as 100% of the light in the aperture (ignoring practical sources of loss; e.g., sampling effects and absorption) can be diffracted into a single pixel (specifically, a diffraction-limited spot.) Also, specially designed and optimized DOE's have been demonstrated that diffract the incident energy into 10-1000 equal intensity spots with efficiencies of 90 % to 100 % [1]. For SLM's that contain from 128^2 to 512^2 pixels it can be seen that for patterns of 1000 spots the SLM pattern can be ~16X to 250X brighter when used as a programmable DOE than when used as an image display.

This result suggests roles for diffractive SLM's in multi-spot scanning and scene illumination that are more general than traditional single-spot mechanical (and also acousto-optic) scanners. Unlike traditional scanners, frame-addressed SLM's are non-inertial and have no memory (as long as the modulating material response time is shorter than the SLM framing time.) Since frame-addressed SLM's can produce sequences of unrelated images, the sequences can be designed that scan multiple spots on arbitrary trajectories and with different velocities and intensities. Because these possibilities are unlike those for previous scanning systems, the goal of this paper is to describe the physical differences between traditional, inertial type scanners and diffractive SLM based illuminators (Section 2), to present the software programming and computational considerations for on line design of the required SLM modulation patterns (Section 3), and to experimentally demonstrate, using a phase-only SLM, some of the generalized scanning and illumination functions that are possible using diffractive, frame-addressed SLM's (Section 4).

While the diffraction patterns that we will present are generated at real-time rates, the SLM modulation patterns usually were designed and optimized off-line. A continuing goal in our research is to develop an attached computer/hardware/software system that automatically designs the modulation patterns at real-time rates [2]. We recognize that any practical system developed will require a tradeoff between optical performance and computational load. This leads to us proposing in Section 3 a three-level, hierarchical design strategy that trades off optical performance versus available computation time. We believe that for many applications it will be reasonable to design patterns automatically at real-time rates and with sufficient optical quality to permit laser designation of multiple moving objects. Herein we describe a proposed multi-spot scanning system in enough detail to permit evaluation of the new functionality provided together with a description of system design considerations and fundamental limitations.

2. Distinctions between inertial scanners and frame-addressed SLM's for scanning

Inertial scanners include galvanometers and acousto-optic Bragg deflectors. These devices have the ability both to point or to continuously scan a laser beam. Even when producing spots at non-sequential locations the beam usually is scanned continuously, which may require the laser to be shuttered if the line between two points is not to be illuminated. Ideally frame-addressed SLM's produce sequential images and shuttering of the laser source is not required. However, stray light can be generated during the transition time between successive SLM frames due to the finite response time of the modulating material. When the SLM framing time is close to the material response time, the energy in stray light can be comparable to the energy in the desired pattern.

Received October 17, 2000; Revised November 25, 2000

Thus at limting frame rates, shuttering of the laser may be necessary.

While inertial scanners can produce continuous line scan automatically, frame-addressed SLM's cannot. Furthermore the time required to scan a line over the full angular range of the device is of the same order as the time to repoint the laser to an arbitrary position. To approximate a continuous scan line with SLM's the diffraction patterns must sample space without leaving any gaps. This problem corresponds to successively sampling the diffraction plane n times with a diffraction limited spot, where n is the number of pixels in a n x n pixel SLM. If the frame rate of the SLM is identical to the frequency of the inertial scanner then the inertial scanner is n times faster than the SLM.

However, this is a worst case comparison which does not take into account the additional capabilities of SLM's to produce multiple spots in parallel and to produce illumination patterns with footprints that are larger than the diffraction limit. The difference in scan speeds between inertial and SLM scanning can be reduced to n/m_1 by using the SLM to scan m_1 diffraction limited spots along the scan direction, or by broadening/blurring a spot by a factor of m_2 which leads to a speed comparison of n/m_2. Also varying combinations of scanning and blurring can be used to obtain a speed comparison of $n/(m_1m_2)$. (Note: This speed comparison is not meant to reflect that both blurring and multiple spots increase scanning more than either does individually. Instead the speed comparison suggests that various combinations of blurring and multiple spots can be used to scan continuously with a SLM.)

Examples of these possibilities either have been presented previously or are presented herein. In Section 4 we present an example of scanning an array of multiple spots. One way to obtain blurring of the spots is to partition the SLM into sub-arrays in which the aperture of each sub-array sets the diffraction limited spot size. Another way is to design an effective apodization into the modulation pattern that blurs or broadens the spots.

We complete this comparison by considering the speeds of various devices. Commercial galvanometric scanners scan at rates up to their resonant frequency ~5 kHz and take small steps at ~0.2 ms [3,4]. Acousto-optic Bragg deflectors can scan in excess of 1 MHz [5,6]. The speeds of the galvanometers are comparable to the frame rate of current liquid crystal on silicon SLM's (10 kHz for the BNS 128 x 128 pixel SLM) [7]. However, for analog phase modulation over a full 2π range nematic liquid crystal is used. Its response time is ~2 to 4 mS which would limit useful frame rates to ~250 Hz. Even at this low rate a single diffraction-limited spot (and also multiple spots) could be continuously scanned over the 128 resolution cells 2 times per second. A faster phase shifting device that is not commercially available at this time is the flexure beam deformable mirror device. Response times of under 10 μS have been reported [8]. Given fast enough frame addressing circuits then ~800 scans per second per spot or faster is possible. Using the frame rate of the BNS addressing circuit as a reasonably practical number gives ~80 scans per second. The SLM used for the experiments in this study is the BNS SLM described above. It is filled with nematic liquid crystal and thus is limited to practical frame rates of around 250 Hz and 2 complete line scans per spot per second.

In order to clarify the differences between frame-addressed SLM's and traditional scanners we have compared their speeds at continuous scanning of the full field of regard. By this comparison SLM's appear to be quite slow. However, the applications we envision for the SLM (while they may require an occasional full field scan) do not require repetitive full field scanning for which traditional scanners are best suited. Specifically, we envision using an SLM illuminator together with feedback from a video camera viewing the illuminated scene to adaptively track and designate objects of interest in the scene. The ability to configure patterns flexibly provides a way to intelligently interrogate the field of regard. Using information gained from previous video frames on object motion provides a way to reduce the total area of the scene that needs to be illuminated. Used in this way it appears likely that even the 250 Hz frame rate SLM reported in this study has adequate frame rate to support adaptive tracking and designation of multiple objects in situations where the motion of the objects is neither too fast nor too erratic to be predicted.

#26795 - $15.00 US
Received October 17, 2000; Revised November 25, 2000

3. Hierarchical DOE design system

In addition to the physical limitations that set SLM frame rate there is also a computational limitation. For the adaptive type applications that we are considering, it usually will not be possible to predetermine an appropriate set of desired diffraction patterns. Therefore it will be necessary to design the required spatial modulations for the SLM on-line as needed.

Current methods of designing fixed pattern DOE's usually use global search and optimization to identify modulation patterns that produce the best possible diffraction patterns. These methods are computationally intensive. The reason is that there are a near infinite number of Fourier transform pairs that have identical intensity diffraction patterns but which have different phase diffraction patterns. In illuminator systems, in which only the intensity of the desired diffraction pattern is of interest, global search and optimization are used to identify the phase diffraction pattern that gives the best performing intensity diffraction pattern. If the SLM can produce any arbitrary complex value of modulation then all diffraction patterns are identical except for an intensity scale factor. The optimization algorithm searches for the modulation pattern that maximizes the intensity (and hence, the diffraction efficiency) of the diffraction pattern. However, most modulators are not fully complex, and thus the diffraction pattern may only approximate the desired diffraction pattern. Then the optimization algorithm needs to trade off diffraction efficiency with accuracy or "fidelity" of the intensity pattern.

Because optimization methods are computationally intensive, it may not be possible to complete the calculation at the frame rate required for a real-time illumination system. Faster, but lower performing, algorithms can be used to reduce computation time. Then, if additional time becomes available (e.g. there is little change in a scene for a period of time) design algorithms that produce higher performance are used. This introduces the hierarchical design strategy needed for on-line design. Fig. 1 summarizes one approach to on-line design that has three levels of computational complexity.

Level 1: Minimum Time Design. The design method starts with a specification of $I_T(f)$ the desired target intensity pattern in the diffraction plane. For a target diffraction pattern composed of an array of spots we may specify the desired $I_T(f_i)$ for the discrete set of frequencies f_i. These values together with any arbitrary set of phases ϕ_i specify the complex light distribution $\mathbf{A}_c'(f_i)$. The complex valued function $\mathbf{a}_c'(x)$ where x is the modulator plane coordinate can then be calculated (the compose function block in Fig. 1a.) Composition can be performed using the inverse fast Fourier transform (IFFT) of $\mathbf{A}_c'(f)$. However from the standpoint of ease of use or numerical efficiency it may be preferred to use known Fourier transform pairs $\mathbf{g}_i(x)\leftrightarrow\mathbf{G}_i(f)$ and $\exp(j2\pi f_i x)\leftrightarrow\delta(f-f_i)$ followed by the superposition of these elementary functions to give

$$\mathbf{a}_c'(x) = \sum_{i=1}^{M} \mathbf{g}_i(x)\exp\left[j\left(2\pi f_i x + \phi_i\right)\right]. \qquad (1)$$

The magnitude of function $\mathbf{a}_c'(x)$ is then rescaled so that the maximum magnitude is unity (normalization step in Fig. 1a with γ initially equal to unity) which produces the normalized function $\mathbf{a}_c(x)$. The normalized function $\mathbf{a}_c(x)$ is mapped through a selected encoding algorithm into a modulation $\mathbf{a}(x)$ that is achievable with the limited modulation range of the non-fully complex SLM. The diffraction pattern intensity $I(f)$ is produced by the optical Fourier transform (OFT).

The normalization step in Fig. 1a is suggested by the fact that SLM's are passive devices. However, the actual reason for this step is that while some algorithms can be applied if magnitudes are greater than unity, many encoding algorithms cannot [2]. Furthermore, of the encoding algorithms that can handle magnitudes that are greater than unity, many of these algorithms nonetheless depend on the exact scaling of the function. (Scaling is discussed further in the next subsection on Level 2 design).

Received October 17, 2000; Revised November 25, 2000

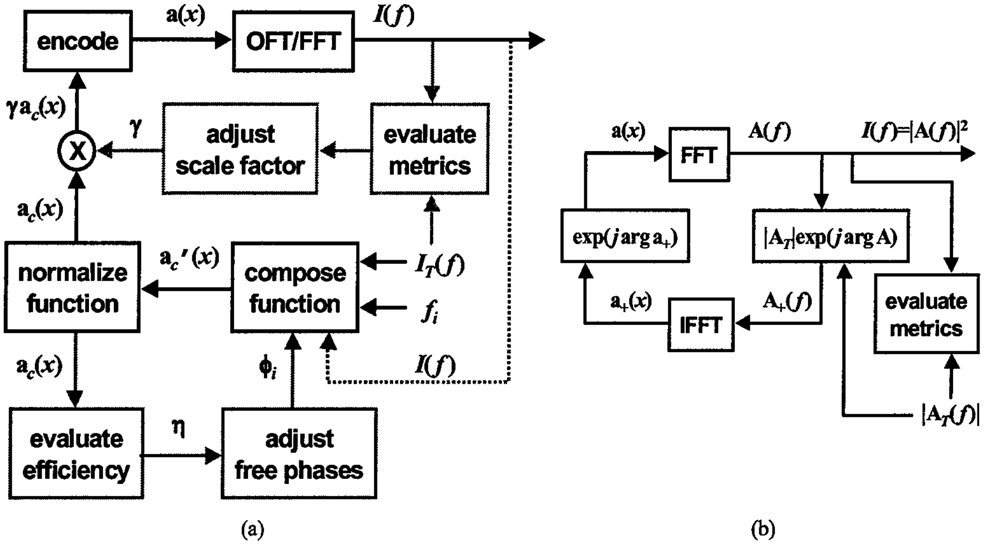


Fig. 1. (a) The proposed hierarchical design system, (b) the iterative Fourier transform algorithm. In (a) Level 1 design is indicated by black lines and boxes. With the addition of the algorithmic portions indicated by red and blue lines, the algorithm becomes a Level 2 or Level 3 design respectively. The errors between the target and resulting intensities (dashed blue line) can be used to compensate the desired function, and this procedure, which is similar to adjustment of the free phases, also is classified as Level 3 design.

The major computational time for the Level 1design includes composing the function $\mathbf{a}_c'(x)$ (which requires when using superposition either $O(NM)$ function calculations for an N pixel SLM and M frequencies of interest f_i, or when using IFFT $O[N\log_2(N)]$ multiplies,) and encoding [which requires $O(N)$ function calculations.] For a small enough number of diffraction spots [i.e. for $M<\alpha\log_2(N)$ where α is a scale factor accounting for the exact differences in computation time between the IFFT and superposition algorithms,] superposition of Fourier transform pairs requires less computation time. While composition of the desired function requires $O(M)$ times more calculations than encoding, the elemental functions can be synthesized in parallel which, at the cost of added (analog or digital) processing hardware, would permit the rate of function calculation to match the rate of encoding.

Level 2: Moderate Time Design. If there is additional computation time available after completing Level 1, the desired complex function can be encoded by multiple algorithms and the encoding that produces the most satisfactory diffraction pattern is selected. Example of this are the hybridization of multiple encoding algorithms [9-11]. The hybridization is characterized by one (or more) parameter(s) γ that defines which regions of the complex plane are encoded by which algorithm. For instance, in the minimum distance-pseudorandom encoding algorithm (MD-PRE) for phase-only SLM's in Ref. [9], values of the scaled function $\gamma\mathbf{a}_c(x)$ (see Fig. 1a) with magnitudes less than unity are encoded by pseudorandom encoding algorithm (PRE) [12,13] and complex values with magnitudes greater than unity are encoded by minimum distance encoding (MDE) [14]. The best performance in terms of specific fidelity measures (e.g. non-uniformity, signal-to-peak-noise ratio) is usually found for a particular blending of MD-PRE as specified by the value of γ. We are unaware of an *a priori* method to select or estimate the best value of γ. This leads to our proposal to adjust blending parameters based on the resulting intensity pattern $I(f)$ (as indicated by the innermost feedback loop in Fig. 1a.) The intensity pattern can be simulated using the FFT, but it can be faster to use the OFT to produce a diffraction pattern and then measure its intensity with a digital camera. Since only intensity, and not phase, is needed to find the optimal value of γ, digital measurement and feedback of the diffraction pattern is reasonably practical.

Level 3: Maximum Time Design. If there is significantly more time available than required to complete Levels 1 and 2, then DOE type design, which includes multiple iterations of a global search and/or optimization method can be employed. The approach indicated in Fig. 1a is to compose and evaluate multiple functions $\mathbf{a}_c(x)$ that satisfy the target intensities $I_T(f_i)$. Generally the higher the diffraction efficiency η of the complex valued function, the more closely $I(f)$ compares to $I_T(f)$ [15].

There are many methods possible to select the phase free parameters ϕ_i (e.g. Monte Carlo selection, gradient search, genetic algorithms and projection onto constraints [15,16]) that could be used in this block of the Fig. 1a flowchart. While $\mathbf{a}_c(x)$ can be computed using the IFFT, this calculation can be avoided by superposition of known Fourier transform pairs, which can be faster, as discussed above. Methods of optimizing a complex function followed by encoding is classified by Mait as an indirect design method [16]. Mait observed that better performance is achieved for direct methods, in which the available modulation values are directly adjusted to produce the best solution, but that this may also entail a greater amount of computation than the indirect method. It is the time constraints in real-time pattern generation that suggest our indirect design strategy that is based on composition of fully complex functions followed by encoding.

One limitation of our indirect procedure is that by first selecting $\mathbf{a}_c(x)$ followed by encoding can introduce deviations between $I(f)$ and $I_T(f)$ [16]. It is possible to compensate values of resulting intensity $I(f_i)$ that are greater (less) than the target intensities I_T (f_i) by reducing (increasing) the magnitude scale of the corresponding functions $\mathbf{g}_i(x)$ in eq. (1) (e.g. using the gradient method of Jared and Ennis [17].) Even with feedback of the resulting values of $I(f_i)$ to adjust $\mathbf{a}_c(x)$ (dashed path in Fig. 1a) this modified procedure, though it has an increased computational load is classified as an indirect design method.

Discussion: We have described a framework for on-line design of modulation patterns for SLM's used in a Fourier transform arrangement. The structure is designed to ensure that at least one realization of the desired diffraction pattern can be implemented in as short a time as possible. There are many algorithms that can be used to implement the various blocks, and we do not restrict the system to specific algorithms. Instead, we focused on the relative computation times of the major blocks in order to provide a first look at key issues limiting design time. A more detailed analysis involving numerical and digital implementations of FFT's, arithmetic and function calculations, depends on many design- and application-specific issues and is beyond the scope of this report. However, we did focus on ways in which the use of the FFT (by OFT) and IFFT (by superposition of known Fourier transform pairs) could be avoided to reduce the computation time. In contrast to Fig.1a, consider Fig. 1b which shows a typical structure used in the fixed pattern DOE design. This is the iterative Fourier transform algorithm (IFTA) [18]. Similar to Fig. 1a, the IFFT can be replaced by using superposition. The FFT in Fig. 1b could be replaced with the OFT; however, in addition to intensity, phase also must be measured, which requires more involved hardware [19] For the proposed framework in Fig. 1a, Level 1 provides the shortest path to an initial design. If even shorter design times than possible with Level 1 are required, then further customization of the design may be possible for some applications. One way is to design for only a limited subset of the N SLM pixels. Examples of this approach are presented in Section 4.

Portions of the proposed structure in Fig. 1a have been demonstrated in off line design studies. In Refs. 9, 11 respectively we have demonstrated a one parameter, and a two parameter Level 2 tuning of blended encoding algorithms. In Ref. 15 Yang *et al.* performed a Level 3 design which attempts to maximize the diffraction efficiency of $\mathbf{a}_c(x)$ by three different search/optimization procedures, followed by a single encoding. The problem is partitioned into separate, non-interacting modules. There is no feedback of the resulting intensity $I(f)$ to fine tune $\mathbf{a}_c(x)$ or the encoding algorithm. These procedures require no evaluations by IFFT. Jared and Ennis used a descent search method similar to the Newton-Raphson method to iteratively correct for changes in the peak intensities of autocorrelations due to the SLM modulation characteristics [17].

Received October 17, 2000; Revised November 25, 2000

Vangular observed that this procedure can be directly adapted to improving the uniformity of the intensity patterns from spot array generators [20]. This method corresponds to the feedback of *I(f)* as indicated by the blue dotted path in Fig. 1a.

4. Illustrative Scanning Sequences

To this point we have presented our conception of a real-time diffractive scanner/illuminator. We have focused on speed limitations due both to framing time of SLM's and computational rate for on-line design systems. These discussions are intended to provide a perspective from which one can appreciate the use of SLM's as real-time scanners. In order to complete this picture we demonstrate unique possibilities for scanning with diffractive SLM's through the presentation of experimentally recorded, live video sequences of the far-field patterns from a phase-only SLM. Specific issues and considerations in using SLM's in this way are brought out by these demonstrations, including opportunities in certain situations for computational speedups.

The demonstrations all are performed with a 120 x 128 pixel (the 8 outermost columns of the device are inoperative), reflective SLM from Boulder Nonlinear Systems, Inc. (BNS). The cell is filled with nematic, parallel aligned liquid crystal. The laser polarization is linear and aligned to a collimated green laser beam (532 nm) to produce phase-only modulation of up to 2π. A lens placed one focal length from a CCD video camera forms a Fourier transform on the imaging area of the camera. Unless otherwise noted, the image of the diffraction pattern covers approximately the entire area between the on-axis (0,0) diffraction order and the (1,1) diffraction order (corresponding to the reciprocal of the pixel spacing). This region of the diffraction plane has 120 x 128 resolution cells or a SBWP of 15,360. There is always a very bright spot on the optical axis which is due to Fresnel reflections from the cover glass and optical losses in the modulating layer [9]. Aside from the on-axis spot the device performs very similar to an ideal phase-only SLM. One other feature of the SLM is the subaperture that defines a pixel. This leads to a sinc^2 intensity rolloff along horizontal and vertical axes of the diffraction plane, and a sinc^4 along the diagonal. In some of the diffraction patterns this rolloff is compensated in the design. Additional details of the experimental apparatus have been reported in Ref. 9.

At this time no complete on-line design system exists. The modulation patterns typically are designed off-line and then loaded in sequence onto the SLM. In some cases, when only Level 1 design is adequate, the patterns are designed and loaded as created. Using a 100 MHz Pentium computer and a C++ program this takes around 0.5 second per frame to compose and encode a 10 spot function. However, no efforts have been taken to optimize the numerical calculations. Specifically we found that the computation time of a 128 x 128 FFT is about 4X faster than the the time used to compose the 10 spot functions. This speed advantage is attributed to the FFT implementation using precomputed values of the complex exponential. However, the purpose of this section, rather them demonstrating speed, is to illustrate various scanning functions and to suggest their usefulness.

Arbitrary scanning. Fig. 2a shows a sequence of five diffraction patterns composed of 63 to 73 spots. The movie shows a repeating sequence of the five images. This illustrates the ability of the SLM to generate rather arbitrary images. Fig. 2b shows the entire area of the image that was captured by the CCD camera. This is nearly the entire diffraction plane between the (0,0) and (1,1) orders. The modulation patterns were designed by the IFTA (see Fig. 1b) followed by MD-PRE of the resulting complex valued function $\mathbf{a}_+(x)$. The performance of the theoretical design for this movie (as well as for all the movies in this paper) is reported in Table 1. It is interesting to compare the performance with $\mathbf{a}(x)$, the traditional IFTA design [which can be interpreted as MDE of $\mathbf{a}_+(x)$.] While IFTA followed by MD-PRE results in a lower diffraction efficiency than traditional IFTA (73 % to 82 %), it has improved fidelity as measured by signal-to-peak ratio (SPR, 12 to 3.5) and nonuniformity (NU, 19 % to 32 %). Using MD-PRE instead of MDE after IFTA, produces a faint background speckle pattern across the entire SBWP (instead of isolated, but more intense noise peaks.) Consequently, even though MDE has a higher signal-to-noise ratio

Received October 17, 2000; Revised November 25, 2000

(SNR) than MD-PRE (720 to 429), the lower average noise of MD-PRE, as measured by SPR, represents the more faithful rendition of the desired diffraction pattern – especially for applications that use the full SBWP of the SLM. These relationships between metrics have been reported in detail our papers listed in *References and links.*

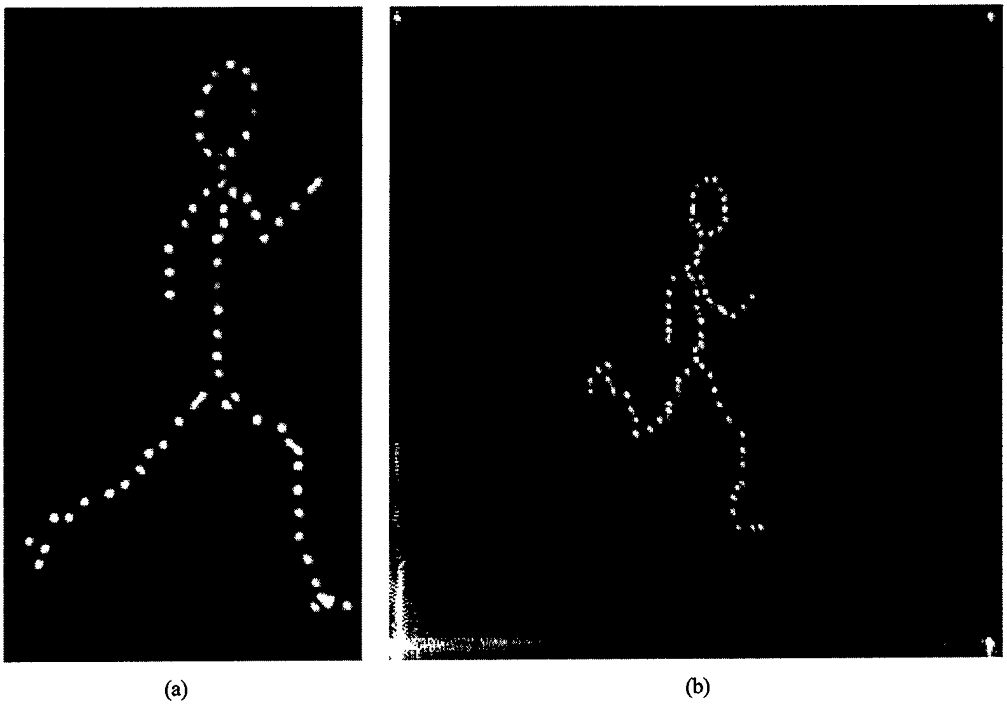

(a) (b)

Fig. 2. (a) (1.62 MB) Movie of arbitrary scanning by SLM, (b) image with (0,0) to (1,1) orders indicated

Table 1. Information on diffraction pattern images, design method and theoretical performance

Figure	Width (%)	Complex function composed by	Encoding method	η (%)	NU (%)	SPR	SNR
Fig. 2	44	IFTA (Fig. 1b)	MD-PRE of $\mathbf{a}_+(x)$	73	19	12	429
Fig. 3a	95	all ϕ_i=0	MD-PRE, γ=1	9	12	4	324
Fig. 3b	95	same as Fig. 3a	MD-PRE, $\gamma=\infty$	11	9	3	404
Fig. 4	95	Based on max. η 1x7 array [9]	MD-PRE, γ=1.4	74	6	72	975
Fig. 5a	55	random ϕ_i, continuous f_i	MD-PRE, γ=1.2	38	10	13	1004
Fig. 5b	55	random ϕ_i, discrete f_i	MD-PRE, γ=1.2	38	10	13	1004
Fig. 6a	100	periodic spatial interleaving	none	100	0	∞	∞
Fig. 6b	100	random spatial interleaving	none	8	17	2	89
Fig. 7a	87	aperture subdivision	MD-PRE, γ=1.5	61	–	–	–
Fig. 7b	87	superposition and subdivision	MD-PRE, γ=1.5	54	–	–	–
Fig. 8a	46	Select ϕ_i randomly	MD-PRE, γ=1.3	25	9	12	111

Theoretical performance: Measured from 128 x 128 FFT of designed function for one selected movie frame.
Width: Width of movie image in x as relative to SBWP in x (i.e. grating frequency). Most images are square.
η: Diffraction efficiency – Ratio of energy at the desired frequencies to energy in the entire diffraction pattern.
NU: Nonuniformity – Standard deviation of the intensity of the spot array relative to the average spot intensity.
SPR: Signal-to-peak-ratio – Average intensity of the spots relative to intensity of the largest noise sidelobe.
SNR: Signal-to-noise-ratio – Average intensity of the spots relative to the average background noise.

Reduction of sidelobes. The problems with noise sidelobes can be better appreciated by comparing the result of encoding the same desired functions by two different encoding algorithms. The particular scanning sequence has five spots moving on five different trajectories. Fig. 3a is the result for MD-PRE and Fig. 3b is the result for using MDE. MDE maps the desired function the closest distance to the available modulator values. This results in a minimum mean squared

Received October 17, 2000; Revised November 25, 2000

error design. However, this systematic method of mapping produces distinct sum and difference frequencies (akin to a hard limiter in communications.) The sidelobes in Fig. 3b are much more intense, and thus, easier to falsely classify as desired signal than the background speckle in Fig. 3a.

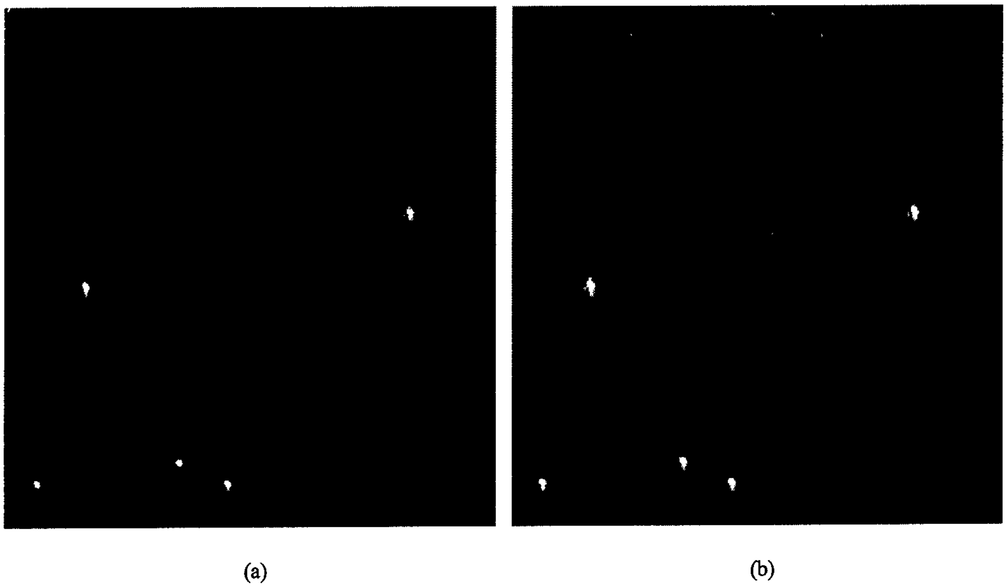


Fig. 3a. Movies of sidelobe generation by (a) MD-PRE (2.41 MB) and by (b) MDE (2.41 MB)

Pattern translation. In some cases it may be desirable to scan a desired pattern to several locations as illustrated in Fig. 4. With a continuous phase, phase-only SLM, scanning of a desired function can be particularly efficient. The computation involves addition of the desired function to a phase ramp, followed by modding the phase, as needed, back into a 2π phase range. Thus while the Fig. 4 image requires the addition of 49 functions to compose the desired modulation, the position can be changed by adding a single function to the encoded function. Note that the design of the desired complex function is based on published maximum diffraction efficiency designs for one dimensional arrays of spots [1]. A numerically efficient method of function composition [O(N) multiplies] is used to compute the one dimension modulation and then form the outer product of the modulation with itself.

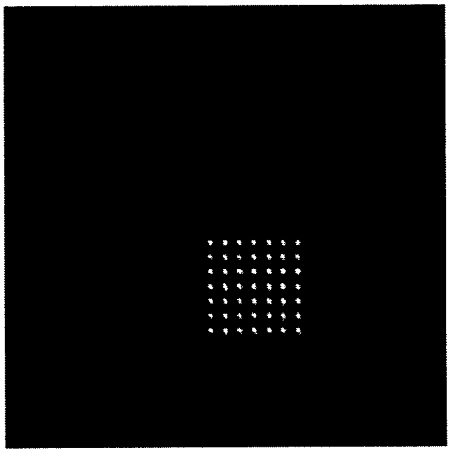


Fig. 4. (1.91 MB) Movie of translation of a fixed pattern.

#26795 - $15.00 US

Received October 17, 2000; Revised November 25, 2000

Continuous scanning. The examples presented so far do not demonstrate continuous scanning. Fig. 5a and the corresponding movie show scanning at various rates from x to y diffraction limited spots diameters per frame. Fig. 5a specifically shows the sum of the multiple frames in the movie. The line rotates 2° per frame and the spots are separated by 6 diffraction limited spot diameters. The fifth line from the pivot point in Fig. 5a corresponds to approximately a one pixel per frame rotation rate. Lines of this radius or less should (and do) appear continuous and lines at greater radii should appear as discrete samples. For the sixth and seventh lines there is still some overlap between the individual spots, which is due to camera saturation and contrast adjustments to the published image. The desired function is composed by addition of complex sinusoids of arbitrary frequencies. A second method of composition is to use the IFFT. Fig. 5b illustrates the result of using a 128 x 128 point IFFT to synthesize a spot pattern. As opposed to Fig. 5a, in Fig. 5b the spots only form at the discrete frequencies corresponding to the frequency sample points of the IFFT. Continuous scanning can be approached by increasing the number of sample points in the IFFT, but at an increasing computational cost. In some cases smaller IFFT's could be used, as well. If the desired modulation is periodic, then one period of the function can be calculated followed by copying or repeating the function to the full size of the SLM. This would be possible for the function used to produce the 7 x 7 spot array in Fig. 4, which consists of a 4 x 4 array of unit cells.

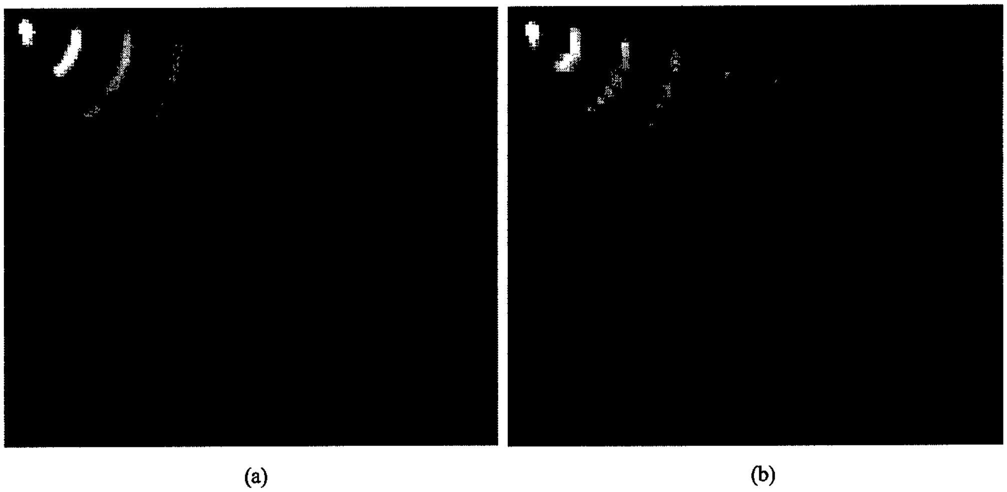


Fig. 5. Movies of continuous scanning (a) by composition (579 KB) and (b) by 128 x 128 IFFT (611 KB). The fixed images in this figure are the summations of the images from the entire sequence.

Replicated scans in parallel. One useful scanning arrangement is to scan the same pattern over multiple fields in parallel. This can be accomplished by under-sampling the SLM, which produces replicas over the full SBWP of the diffraction plane. An example of this is shown in Fig. 6a. Here a sample of the desired function is programmed every fourth pixel in *x* and *y*. In this example 16 linear phase ramps are spatially multiplexed to produce 16 spots replicated into 16 regions. Because each elementary function is phase-only and spatially orthogonal this modulation function has a theoretical diffraction efficiency of 100%. With 256 spots the entire 120 x 128 SBWP of the diffraction plane could be systematically scanned in 60 frames. This corresponds to an equivalent raster scan rate (using a single spot scanner) that is twice the SLM frame rate.

A simple way to produce a nearly identical pattern without replications is to randomly sample, rather than regularly sample the desired function. This method was originally reported by Davis and Cottrell [21]. The method does generate speckle. An example of this design method is presented in Fig. 6b. A speckle pattern background is produced as a result of the random spatial multiplexing of the multiple functions. Also the intensity of individual spots can be adjusted by

#26795 - $15.00 US
Received October 17, 2000; Revised November 25, 2000

increasing or decreasing the percentage of the time that the corresponding modulation function is sampled.

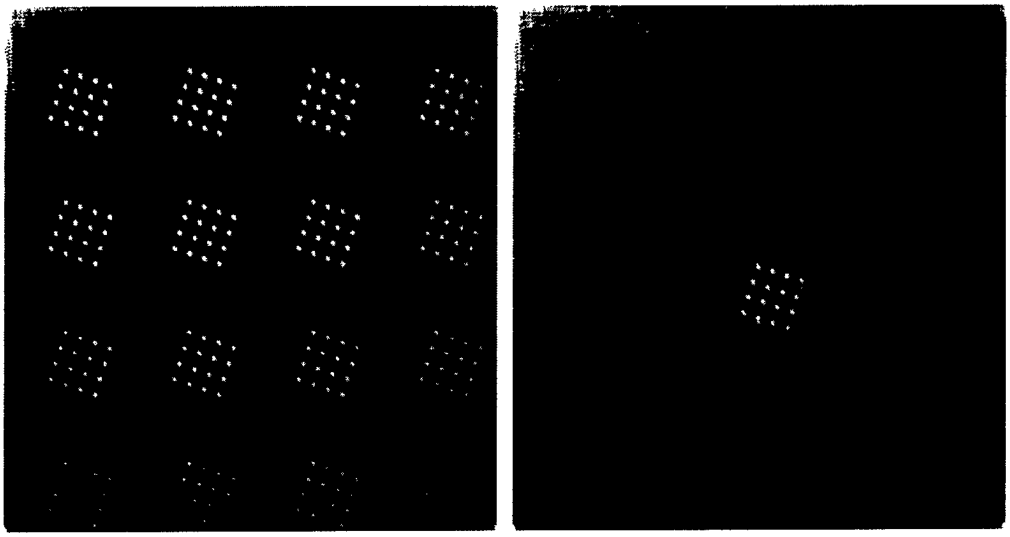


Fig. 6. Movie of (a) parallel replicated scanning by periodic sampling (2.13 MB) and (b) by non-replicated scanning by random sampling (2.23 MB) of the elementary functions.

Variable resolution scanning. Another method of increasing scanning speed (mentioned in Sec. 2) is to design spots that have greater widths than the diffraction limited spot width of the SLM aperture. For the result shown in Fig. 7a, the desired modulation plane function is divided into 9 square regions (32 x 32 pixels each), 2 rectangular regions (60 x 32 pixels) and 1 rectangular region (24 x 96 pixels). Each region is filled with a circular or elliptical aperture function of unity transmittance. The surrounding region of zero transmittance is encoded by MD-PRE. The resulting Airy diffraction patterns are desirable because the first diffraction ring is considerably lower in intensity than the sidelobe of sinc^2 patterns. In Fig. 7b a second layer of

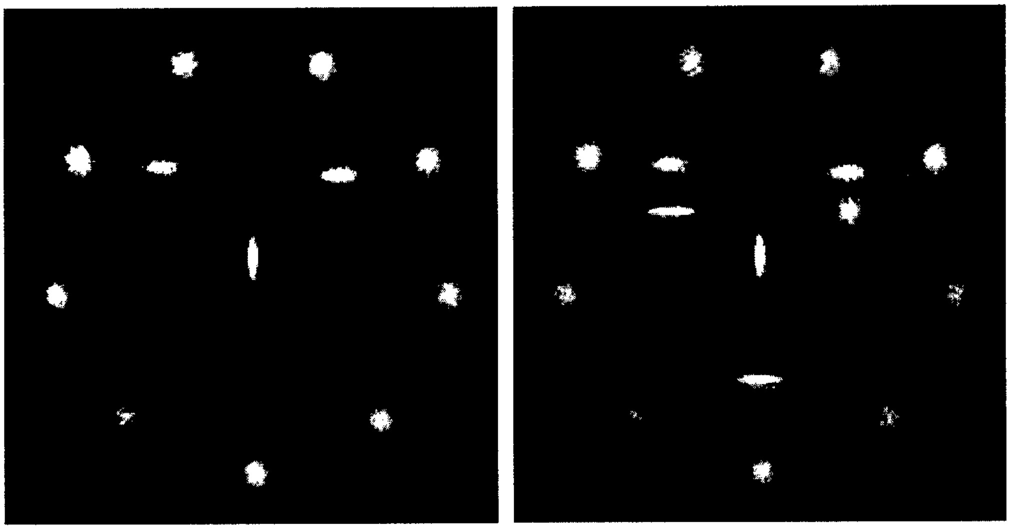


Fig. 7. Movie of multiple widened spots (a) by parallel division of the SLM into multiple SLM's (578 KB) and (b) by adding two sets or layers of spatially multiplexed functions together (1.12 MB.)

Received October 17, 2000; Revised November 25, 2000

elliptically windowed functions is added to the first layer of functions to produce the three additional spots. Two of the bounding rectangles are 120 x 20 pixels and one is 30 x 30 pixels.

Time-averaged scanning. Statistically based encoding algorithms [9-13] necessarily produce noise which is manifested as deviations of the actual diffraction pattern intensities from the desired intensities and speckle background. As seen in Figs. 3-7, these noise effects can be kept to manageable levels for many practical designs. However, these effects can be reduced further by ensemble averaging. Specifically, statistical encoding algorithms produce the desired intensity pattern on-average plus a background that corresponds to the average speckle intensity [12]. Multiple realizations are produced by performing repeated encodings of the desired modulation function. For each encoding a unique random sequence is used. The resulting diffraction plane intensity patterns are averaged together. An experimental demonstration of this procedure for an encoding of an apodized aperture is presented in Ref. 12. Fig. 8 illustrates the improvement in performance of the 7 x 7 spot array design of Fig. 4 for 50 realizations used in the average. The computational load is O[*N*] function calculations for each encoding and O[*N*] additions to add the new intensity pattern to the ensemble. Therefore, if the pattern is not changing over multiple cycles, then a time-averaging procedure can be used to produce a more accurate realization of the desired diffraction pattern, with minimal increase in the computational requirements.

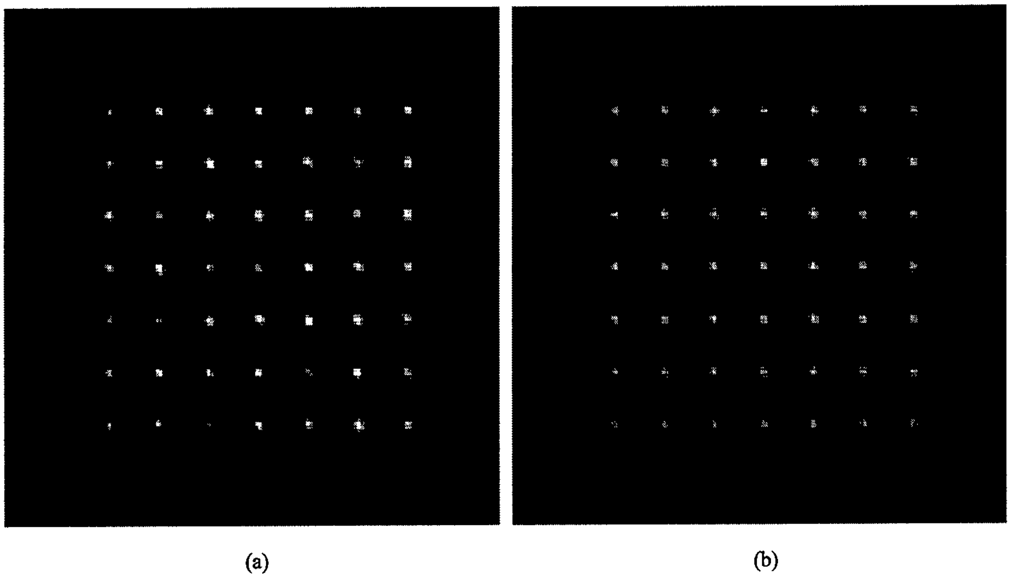
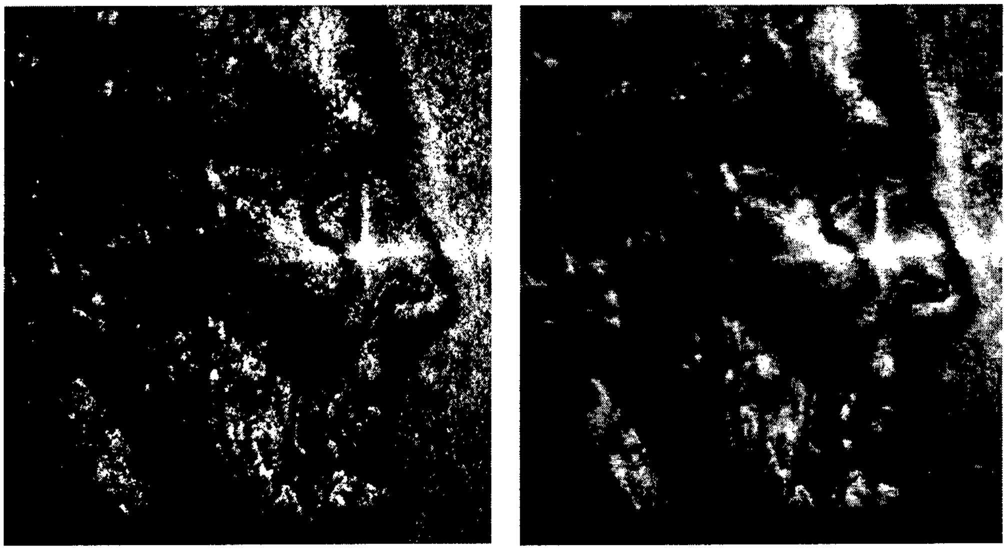


Fig. 8. Image of (a) individual realizations of the spot array (337 KB) and (b) average result of 50 individual realizations of the spot array.

Broad area scene illumination. Modulation patterns can be specified that uniformly illuminate the diffraction plane. As pointed out in Section 1, there is no energy advantage to using diffractive SLM's for this purpose. However, the addition of a broad illumination capability to a multi-spot pattern generator could prove useful for adaptive tracking and designation of objects. While, it should be possible to designate objects and update their positions based on changes in the intensity of the spots reflected from objects, a broad area search is required initially to identify the objects of interest. Using the SLM to illuminate the scene reduces the amount of hardware needed, since a single video camera can be used to observe the entire scene and to monitor the intensity of laser spots reflected from the objects in the scene. Laser illumination is not only required for use in the dark, but it also is useful for lighted scenes. In lighted environment, a narrowband color filter would be placed over the camera to remove other contributions of light and to maximize the detectability of the laser returns. In some situations the laser illumination even can be weak enough to go unnoticed by an observer in the scene.

To produce broad area illumination we programmed the SLM with a sequence of random phases that are uniformly distributed between 0 and 2π. This diffuser produces a speckle pattern at the Fourier transform plane on a copper coin. The coin is illuminated and imaged at ~20° from vertical (Fig. 9a.) The illumination quality can be brought closer to that of incoherent light by averaging multiple realizations of the diffuser. Fig. 9b shows the average image for illumination with 50 statistically independent diffusers.



(a)

(b)

Fig. 9. Image of (a) individual realization of a speckle-illuminated coin and (b) average result of 50 individual realizations of the speckle-illuminated coin.

5. Summary and Conclusion

Using a frame-addressed SLM in a Fourier transform arrangement provides a variety of intensity patterns that may provide much more general scanning capabilities than is possible with traditional inertial type scanners. This increased flexibility comes at a cost of an increased computational load. However, if the system is used as part of a multi-object tracking system or in a vision-guided robotic navigation system, then the computational load associated with on-line design of the diffraction patterns may be commensurate with the loads associated with image processing and supervision of the entire system. In light of the available response time of an adaptive system, a general approach to on-line diffraction pattern has been presented that emphasizes producing designs of increasing fidelity as more computation time is available. The general approach requires iterative optimization or global search if adequate time is available. However, specialized modulation patterns can be devised using spatial multiplexing (as in Figs. 6 and 7) that do not require either optimization or encoding. Other situations can be exploited for speedups, such as translating an identical pattern (in Fig. 4) or imposing special structure on the desired function (e.g. rectangular separability in the design of the spot array in Fig. 4 [13]). Many other possibilities might be exploited by taking into account the scene environment (e.g. spatial extent, velocity and number of objects in the scene.)

Based on the above considerations we believe that affordable scanning systems could be developed that run at practically useful rates. These systems could provide powerful advantages to advanced and intelligent military and commercial systems. The SLM scanners could be used to convert a single laser target designator into a multi-target designator. It could be applied to long range designation of objects in outer space, to shorter range designation on the battlefield,

Received October 17, 2000; Revised November 25, 2000

and to close-in weapons defense of naval vessels. For autonomous control a vision system with multi-object tracking is needed to steer, point and adapt the laser spots. A commercial application of the same system is to coordinate the activities of several distributed robotic package handlers. Lasers used to highlight the packages would make their motion easier to track with machine vision. A laser pattern generator and vision system mounted on a mobile robot could support feature-based navigation. The further development of a diffractive SLM based scanner would be interesting and challenging in terms of algorithm refinement and integration.

Acknowledgments

The funding for this study was from National Aeronautics and Space Administration cooperative agreement NCC5-222, Office of Naval Research grant N00014-96-1-1296, and Ballistic Missile Defense Organization (through Air Force Research Laboratory) grant F19628-99-C-0084.

Received October 17, 2000; Revised November 25, 2000

Improved-fidelity error diffusion through blending with pseudorandom encoding

Li Ge, Markus Duelli,* and Robert W. Cohn

The ElectroOptics Research Institute, University of Louisville, Louisville, Kentucky 40292

Received December 17, 1999; revised manuscript received May 8, 2000; accepted May 17, 2000

Error diffusion (ED) and pseudorandom encoding (PRE) methods of designing Fourier transform holograms are compared in terms of their properties and the optical performance of the resulting far-field diffraction patterns. Although both methods produce a diffuse noise pattern due to the error between the desired fully complex pattern and the encoded modulation, the PRE errors reconstruct uniformly over the nonredundant bandwidth of the discrete-pixel spatial light modulator, while the ED errors reconstruct outside the window of the designed diffraction pattern. Combining the two encoding methods produces higher-fidelity diffraction patterns than either method produces individually. For some designs the fidelity of the ED–PRE algorithm is even higher over the entire nonredundant bandwidth than for the previously reported [J. Opt. Soc. Am. A **16**, 2425 (1999)] minimum-distance-PRE algorithm. © 2000 Optical Society of America [S0740-3232(00)00609-8]

OCIS codes: 230.6120, 090.1760, 070.2580, 030.6600.

1. INTRODUCTION

This paper continues an ongoing study on the development of new procedures for designing Fourier transform holograms.[1–8] The focus of the study has been to develop algorithms that can be computed in real or near-real time and that demonstrate good (rather than optimal) optical performance within the available time constraints for practical spatial light modulators (SLM's). These constraints frequently include that the SLM represents only a limited range of complex values (e.g., phase-only, quantized phase-only, coupled amplitude-phase) and that the SLM has a relatively small number of pixels [or equivalently, space–bandwidth product (SBWP)] compared with fixed-pattern diffractive optical elements and holograms. The motivation behind this research is the development of programmable optoelectronic systems that can automatically design and implement Fourier transform holograms in response to the unanticipated data presented by real-world situations.[9] Examples of such proposed systems include multispot-laser beam-steering systems, multitarget laser designator systems, and distortion-invariant pattern recognition systems implemented with composite filters in a coherent optical correlator.[9]

On the basis of these constraints of the SLM and of the real-world environment, we have avoided computationally intensive global searches for an optimal modulation, and instead we have considered both (1) noniterative encoding of the desired complex-valued function into a spatial modulation pattern and (2) iterative encoding in which one, or a few, free parameter(s) are varied to produce a suboptimal design.

Single-pixel encoding[8] (rather than group-oriented encoding)[10,11] is used. It maps each desired complex value into the modulation value of a corresponding single pixel, without consideration of the modulation values of the surrounding pixels. Single-pixel encoding has the advantages of low computational overhead, and the encoded modulation can have a SBWP that is identical to the SBWP of the desired (spatially sampled) signal. Since the SBWP of the SLM is identical to the number of pixels in the SLM, the far-field diffraction pattern can be reconstructed anywhere within the nonredundant bandwidth (NRB), i.e., the reciprocal of the pixel pitch. Group-oriented methods produce unwanted diffraction patterns within the NRB, which reduce the usable bandwidth to less than the SBWP of the SLM. Examples of single-pixel encoding include

1. Minimum-distance encoding[2] (MDE), in which each desired complex value is mapped to the closest modulation value produced by the SLM. For continuous-phase, phase-only SLM's this corresponds to the classical kinoform[13] or, in the case of pattern recognition filters, the phase-only matched filter.[14]
2. Pseudorandom encoding[1,4–6] (PRE), in which multiple modulation values are randomly selected on a percentage basis so that the expected value of the modulation equals the desired complex value.
3. Blended methods (referred to as MD–PRE)[2,3,8] that encode some desired values by MDE and other values by PRE.

Error diffusion[15–20] (ED) is yet another way to encode complex-valued functions in much less computation time than is required with global search methods. As with single-pixel methods, each desired complex value is encoded as a modulation value of a corresponding SLM pixel. However, the modulation value is also determined by the encoding errors (the difference between a desired value and an encoded value) of a few nearby pixels that have previously been encoded. The weighting factors for encoding errors are chosen so that the Fourier transform of the encoding errors is spatially separated from the desired diffraction pattern. Thus, as with group-oriented encoding, the ED reconstruction is limited to a bandwidth that is a fraction of the NRB of the SLM. This limitation does not restrict the location of the reconstruction within the NRB, since the error weighting factors can be changed for each ED design to maintain separation between the error and the desired reconstruction.[18] However, the design of such weightings can be numerically complex or even impractical for desired patterns that span the NRB (e.g., a pattern of a few widely separated, randomly located spots.)

For wide-bandwidth diffraction patterns it may be simpler to use PRE methods for which the encoding errors reconstruct as a uniform-level noise pattern over the entire NRB. Thus, rather than using spatial separation between the desired pattern and the error pattern to obtain good performance, PRE attempts to distribute the error energy over the entire NRB, which results in low average error intensity everywhere. However, the maximum-intensity noise peak can be on the order of 10× larger than the average noise intensity. This is a consequence of the error pattern statistics, which are identical to the statistics of laser speckle;[1] specifically, the error pattern intensities are exponentially distributed, which makes possible a few noise peaks that are much more intense than the average. MDE also tends to produce a few bright noise spikes that appear at sum and difference frequencies of the desired pattern. Particular blendings of MDE with PRE have been demonstrated to produce lower peak noise (and also more accurate approximation of the intensities of the desired diffraction pattern) than either method individually.[3,8,21] We note that the reduction of noise spikes, particularly in spot array generators designed with a minimum-distance criterion, has been a motivation both for ED by Kirk *et al.*[19] and for MD–PRE.[3,8]

In this paper we present, to our knowledge for the first time, comparisons of ED and PRE in terms of their properties and optical performance. The general differences that we discussed above are brought out further by application of each encoding method to the same desired fully complex function and comparison of the resulting diffraction patterns. In addition to reviewing the ED and PRE algorithms, we show that a hybrid algorithm can be constructed out of the individual ED and PRE algorithms. We show that the ED–PRE blended algorithm outperforms both ED and PRE in terms of two fidelity metrics that measure noise spikes and accuracy between the desired and the resulting diffraction pattern. We also include comparisons of the blended ED–PRE method with the earlier MD–PRE method.

Section 2 presents a mathematical description of each algorithm evaluated in the study. Their performance is evaluated by computer simulation. The simulation procedure is described in Section 3, and the results and the performance comparisons are presented in Section 4.

2. DESCRIPTION OF THE ENCODING ALGORITHMS

The algorithms presented in this paper are specialized for phase-only SLM's that produce any value of phase continuously over 360°. Encoding algorithms for many other modulation characteristics are possible. A few of these include encoding for amplitude-phase coupled,[3,5,12] binary quantized,[16] and *m*-ary quantized characteristics.[6,8,12,17,21] Although specifying and evaluating such a variety of algorithms is beyond the scope of this paper, this section would also help in the development of blended algorithms for modulation characteristics other than phase-only. We first present the three individual algorithms MDE, ED, and PRE, followed by blended MD–PRE and the new ED–PRE.

When algorithms are encoded a desired modulation is specified. The desired function will be considered to be a discretely sampled two-dimensional array of, in general, complex numbers. The x coordinate will be associated with the index i and the y coordinate with index j. For the simulations in this paper the desired function and the SLM pixels always are equally spaced. A fully complex value from the desired function is written as $\mathbf{a}_{c\,ij}$ where boldface indicates a complex-valued quantity. After encoding, the function is mapped into the SLM modulation values $\mathbf{a}_{ij}$.

A. Minimum-Distance Encoding

MDE is presented because (1) it is a limiting case of ED when all the nearby error weighting coefficients are set to zero, (2) it is part of the blended MD–PRE algorithm, and (3) for phase-only SLM's it has the interesting property of producing the maximum diffraction efficiency for any encoding of a given complex-valued function.[22] An equivalent statement is that MDE produces the smallest total encoding error.

MDE is illustrated in Fig. 1(a). The unit circle represents the complex modulation characteristic of a phase-only SLM. The encoding algorithm is a direct point-by-point mapping of the desired value (along radial lines) to the closest point on the modulation characteristic, which is identical to kinoform design. The MDE algorithm for continuous phase-only modulation is written

$$\mathbf{a}_{ij} = \exp[j \arg(\mathbf{a}_{c\,ij})], \tag{1}$$

which has error

$$\mathbf{e}_{ij} = \mathbf{a}_{c\,ij} - \mathbf{a}_{ij}. \tag{2}$$

Desired values of any magnitude $[0,\infty]$ are mapped to the unit circle, as illustrated in Fig. 1(a).

B. Error Diffusion

ED [Fig. 1(b)] can be viewed as a modified version of MDE in that the value

$$\mathbf{a}_{ij} = \exp[j \arg(\mathbf{b}_{ij})], \tag{3}$$

where

$$\mathbf{b}_{ij} = \mathbf{a}_{c\,ij} + (\varepsilon_{i-1,j} + \varepsilon_{i,j-1})/2, \tag{4}$$

is a perturbed value of $\mathbf{a}_{c\,ij}$ and where the perturbations are the errors

$$\varepsilon_{ij} = \mathbf{b}_{ij} - \mathbf{a}_{ij} \tag{5}$$

for two nearest-neighbor samples to $\mathbf{a}_{c\,ij}$. The specific error samples and their weights (1/2 each) are identical to those used by Weissbach *et al.*[17] Other combinations of weights and error samples have been used to vary the error reconstruction pattern.[18] The specific algorithm given here will be used in the computer simulations that follow.

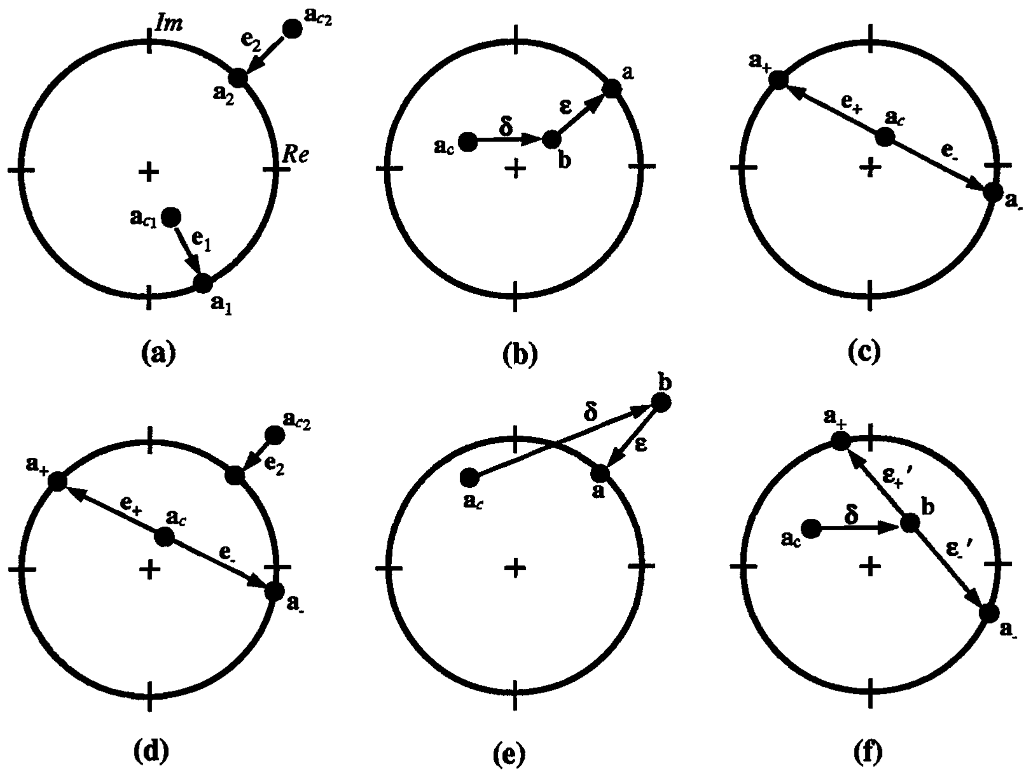


Fig. 1. Illustration of the encoding methods: (a) MDE, (b) ED, (c) PRE, (d) MD–PRE, (e) ED–PRE for $\mathbf{b}_{ij}$ outside the unit circle, (f) ED–PRE for $\mathbf{b}_{ij}$ inside the unit circle. The error diffused forward from previously encoded pixels is represented in the illustration by $\delta_{ij} \equiv (\varepsilon_{i-1,j} + \varepsilon_{i,j-1})/2$. In (f) the encoding error $\varepsilon' \equiv \varepsilon/\chi$, where ε is the amount of the encoding error that is diffused forward with use of Eq. (4).

C. Pseudorandom Encoding

For each desired complex value $\mathbf{a}_{c\,ij}$, PRE prescribes a random (complex-valued) variable $\mathbf{a}_{ij}$ such that the expected value $\langle \mathbf{a}_{ij} \rangle = \mathbf{a}_{c\,ij}$. One way that this condition can be met for phase-only SLM's is illustrated in Fig. 1(c). Here the encoded phase is $\psi_{ij} \equiv \arg(\mathbf{a}_{ij}) = \psi_{c\,ij} \pm \nu_{ij}/2$, where $\psi_{c\,ij} \equiv \arg(\mathbf{a}_{c\,ij})$ is the desired phase and $\pm\nu_{ij}/2$ is a phase offset. The positive or negative sign is randomly selected, each with a probability of 50%. If the uniformly distributed random variable $s_{ij} \in [-0.5, 0.5]$ is used, then the phase can be encoded as $\psi_{ij} = \psi_{c\,ij} + \mathrm{sgn}(s_{ij})\nu_{ij}/2$. For this phase random variable the expected value of complex transmittance is

$$\langle \mathbf{a}_{ij} \rangle = \cos(\nu_{ij}/2)\exp(\psi_{c\,ij}). \qquad (6)$$

Values of the desired magnitude $|\mathbf{a}_{c\,ij}| \in [0, 1]$ are then encoded by choosing values of $\nu_{ij} \in [0, \pi]$ such that

$$\nu_{ij} = 2 \arccos(|\mathbf{a}_{c\,ij}|). \qquad (7)$$

We will refer to this specific PRE algorithm as the inverse cosine algorithm.[5]

The error for encoding a single desired value is calculated by Eq. (2) with the encoded complex modulation $\mathbf{a}_{ij}$ as calculated in this subsection. This definition of encoding error differs from that in our previous papers.[1–8] In the earlier studies the *average* error contribution resulting from encoding a single pixel was evaluated rather than the *actual* contribution. This distinction is important because while several PRE algorithms have been developed for phase-only SLM's, they produce identical average errors and diffraction patterns of essentially identical performance.[7] However, when these various algorithms are blended with ED, (1) the actual error contributions are error diffused and (2) the performance of the diffraction patterns differs depending on which PRE algorithm is used.

For the particular desired functions $\mathbf{a}_{c\,ij}$ that we considered in this study, the inverse cosine PRE algorithm produces somewhat higher-fidelity reconstructions than the inverse sinc PRE method[1] and the phase reversal PRE method described in Sec. 3.C of Ref. 8. Therefore detailed simulations made with these alternate PRE algorithms are omitted because they provide little additional information over the results (presented in Section 4) found with the inverse cosine algorithm.

D. Blended Minimum-Distance Pseudorandom Encoding

MD–PRE was first introduced in Ref. 2 and has been further developed in Refs. 3 and 8. This method trades off desirable performance properties of the MDE algorithm with those of the PRE algorithm. MDE applied to a phase-only SLM is known to produce the highest possible diffraction efficiency for a given function of any encoding algorithm.[22] However, MDE is quite susceptible to intermodulation distortion[23] and (especially for spot array generators) can produce large sidelobes at sum and difference frequencies of the desired diffraction pattern.[24] PRE results in lower diffraction efficiency, and the errors between the desired and the resulting design are due to interference with background speckle, which is a necessary by-product of the PRE method. Various simulations and experiments have shown that blending leads to overall better performance.[2,3,8] Specifically, the background speckle intensity is reduced and the efficiency is increased over PRE alone, and the distortion and sidelobe levels are reduced over MDE alone.

The MD–PRE algorithm can be expressed as follows:

If $|\mathbf{a}_{c\,ij}| > 1$ encode by MDE algorithm [Eq. (1)] as illustrated in Fig. 1(a).

Otherwise encode by PRE algorithm in Subsection 2.C as illustrated in Fig. 1(c).

The combined algorithm is illustrated in Fig. 1(d).

The performance of the algorithm depends on a single free parameter γ. This parameter is the maximum magnitude of the desired complex values $\mathbf{a}_{c\,ij}$. With PRE it is possible to encode only values of magnitude less than or equal to unity.[7] Therefore for PRE alone we normally scale the complex values of the desired function so that γ equals unity. PRE also permits encoding for values of γ that are less than zero, but this produces increased levels of speckle noise and lower diffraction efficiency.[4] However, MDE can encode values of any magnitude. For MD–PRE we have always found a particular value of γ greater than unity that minimizes the approximation errors and another value that minimizes the maximum-intensity noise sidelobe in the diffraction pattern. Currently there is no method that provides an *a priori* estimate of the optimal value of γ. Instead, the best value of γ is found by repetitive simulations of the encoding algorithm. We will show in Section 4 that the free parameter γ also controls the performance of ED and ED–PRE algorithms.

E. Blended Error-Diffusion Pseudorandom Encoding

The blending of ED with PRE is similar in philosophy to the blending of MDE with PRE. However, there are multiple possible ways that this might be accomplished. Among the various blending approaches we have considered are the following:

I. Apply ED [Eq. (3)] to values of $\mathbf{b}_{ij}$ outside the unit circle and apply PRE to values of $\mathbf{b}_{ij}$ inside the unit circle. The error from Eq. (5) for values encoded by ED is diffused forward by use of Eq. (4). The error from Eq. (2) for values encoded by PRE is not diffused forward; i.e., it is treated as zero in Eq. (4). The rationale for this is that the average error produced by PRE is automatically diffracted into speckle background and does not also need to be diffused into adjacent pixels.

II. Same as I except that the average error from PRE is diffused forward by Eq. (4).

III. Same as I except that the actual error from PRE is diffused forward by Eq. (4).

IV. Same as I except that a fraction $\chi \in [0,1]$ of the actual error from each PRE encoded value is diffused forward by Eq. (4). We will define the amount of error that is diffused from a PRE encoded pixel (rather than the total error) as

$$\varepsilon_{ij} = \chi(\mathbf{b}_{ij} - \mathbf{a}_{ij}). \qquad (8)$$

We empirically found through various simulations and experimentation that method IV (for optimized values of χ) produces significantly better performance than methods I and II and somewhat better performance than method III. The sensitivity of method IV with respect to values of $\chi \in [0.5,1]$ is apparent though not dramatic (as is illustrated in Section 4.) Therefore the performance of method IV depends on the two free parameters γ and χ. Since PRE, MD–PRE, and ED–PRE all use random variables, the performance of each algorithm also depends on the particular random sequence used. In Section 4 we also present the variations in performance for an ensemble of random sequences. We are not recommending that ED–PRE be optimized in terms of all three variables (γ, χ, and which sample of an ensemble of random sequences is selected) for our envisioned real-time and near-real-time applications, but rather we present these analyses to provide insight into the performance of the algorithms.

In the remainder of this paper, method IV will be referred to as the ED–PRE algorithm. It is implemented as follows:

Given the $N = nm$ desired values $\mathbf{a}_{c\,ij}$, N uniformly distributed random numbers $s_{ij} \in [-0.5, 0.5]$, and specific values for γ and χ,

1. Normalize all $N\mathbf{a}_{c\,ij}$ so that the maximum of the values $|\mathbf{a}_{c\,ij}|$ equals γ.
2. For $i = 1$ to n and $j = 1$ to m.
3. Calculate $\mathbf{b}_{ij}$ using Eq. (4).
4. If $|\mathbf{b}_{ij}| > 1$ Encode $\mathbf{b}_{ij}$ to $\mathbf{a}_{ij}$ using Eq. (3) Calculate encoding error ε_{ij} using Eq. (5). Otherwise Encode

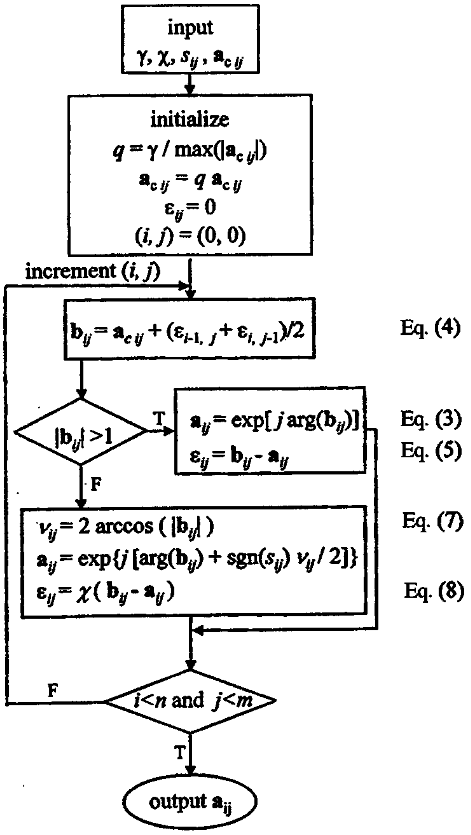

input
$\gamma, \chi, s_{ij}, \mathbf{a}_{c\,ij}$
initialize
$q = \gamma / \max(|\mathbf{a}_{c\,ij}|)$
$\mathbf{a}_{c\,ij} = q\, \mathbf{a}_{c\,ij}$
$\varepsilon_{ij} = 0$
$(i, j) = (0, 0)$
increment (i, j)
$\mathbf{b}_{ij} = \mathbf{a}_{c\,ij} + (\varepsilon_{i-1,\,j} + \varepsilon_{i,\,j-1})/2$
Eq. (4)
$|\mathbf{b}_{ij}| > 1$
T
$\mathbf{a}_{ij} = \exp[j \arg(\mathbf{b}_{ij})]$
$\varepsilon_{ij} = \mathbf{b}_{ij} - \mathbf{a}_{ij}$
Eq. (3)
Eq. (5)
F
$v_{ij} = 2 \arccos(|\mathbf{b}_{ij}|)$
$\mathbf{a}_{ij} = \exp\{j[\arg(\mathbf{b}_{ij}) + \mathrm{sgn}(s_{ij})\, v_{ij}/2]\}$
$\varepsilon_{ij} = \chi(\mathbf{b}_{ij} - \mathbf{a}_{ij})$
Eq. (7)
Eq. (8)
F
$i<n$ and $j<m$
T
output $\mathbf{a}_{ij}$

Fig. 2. Flow chart for the ED–PRE algorithm.

$\mathbf{b}_{ij}$ to $\mathbf{a}_{ij}$ using the PRE method in Subsection 2.C and Calculate ε_{ij}, the encoding error to diffuse forward, using Eq. (8).

A detailed flow chart of the complete ED–PRE algorithm is presented in Fig. 2. Figures 1(e) and 1(f) graphically illustrate the blended encoding process for values of $\mathbf{b}_{ij}$ both inside and outside the unit circle.

3. SIMULATION PROCEDURES

This section presents simulated diffraction patterns for, and evaluates the performance of, the various encoding algorithms for two specific test patterns, and the procedures used in those experiments, are described.

The ED, MD–PRE, and ED–PRE, as described in Section 2, are implemented for values of $\gamma \in [1, 2.5]$ in increments of 0.1. PRE corresponds to MD–PRE with $\gamma = 1$. Also, the performance of MDE is reported. It corresponds to MD–PRE with $\gamma = \infty$. It was observed for the functions encoded in this study for $\gamma \geqslant 2.5$ that the performance of MD–PRE is nearly identical to the performance of MDE. ED–PRE is implemented for values of $\chi \in [0, 1]$ in increments of 0.1. One set of simulations is performed, with the identical set of random numbers s_{ij} used for each encoding and for each value of γ and χ. These simulations are used to make direct comparisons of the performance of the algorithms. A second set of simulations is performed to determine the statistical variations in an ensemble of runs. For this evaluation the simulation is repeated an additional 20 times, each time with a different N sample sequence s_{ij}. The best-case and worst-case performance is reported for the 21 trials of each algorithm over a range of values of γ and for fixed values of χ.

Two test functions are selected for encoding. For continuity with our previous studies we use the same $N = 128 \times 128$ pixel test function $\mathbf{a}_{c\,ij}$, which produces a 7×7 spot array.[6,8] This function reconstructs off axis. The function was selected because it is typical of current diffractive optic designs that have diffraction efficiencies close to the theoretical maximum.[22] Specifically, when the fully complex function is encoded by MDE, the diffraction efficiency η is found to be 96% (see Table 1 and 2 below). The fully complex function itself has a diffraction efficiency of ~44%. The efficiency indicates how much amplitude information is in the function. Since diffraction efficiency of the fully complex function $\mathbf{a}_{c\,ij}$ can be shown to be identical to the average intensity of the modulation,[4] the square root of η gives the root-mean-square amplitude of the complex function of 0.66. This shows that a significant part of the fully complex function is not phase-only and requires some type of encoding. For comparison, the diffraction efficiency of a fully complex function used to generate a 49-spot array can be as low as 2% (when the phases of all diffracted spots are identical), and the root-mean-square amplitude is then 0.14. We know that the encoding errors [see Eqs. (2), (5), and (8)] would be greater and the performance would decrease for the lower-efficiency complex function;[1,4] however, further consideration of this point is beyond the scope of this paper.

The second test function is identical to the first except that a linear phase ramp has been removed from the first test function, so that its diffraction pattern reconstructs centered on the optical axis. Diffraction patterns centered at multiple locations are evaluated to give a more complete appreciation of the performance of the various encoding algorithms. (However, in practice one should anticipate the presence of an on-axis order because of practical limitations in perfectly controlling the fabrication of a diffractive optic or the phase settings of a SLM.)

The diffraction pattern is simulated by performing a fast Fourier transform (FFT) of the test function. The modulation value of an SLM pixel is represented by a single complex number. The 128×128 array of numbers is zero padded to form a 512×512 array that is transformed by the FFT. The padding samples and interpolates the diffraction pattern at 1/4 the diffraction limit, thus producing a realistic-looking diffraction pattern.

Diffraction patterns simulated in this way are evaluated to determine signal-to-noise ratio (SNR), signal-to-peak-noise ratio (SPR), and nonuniformity (NU). NU measures the relative deviation of the spot array from perfectly uniform. The peak intensities of the 49 spots is measured. The average and standard deviation of the intensities are calculated, and NU is the ratio of the standard deviation to the average intensity.

The average intensity of the 49 spots is also used in SNR and SPR calculations. For SNR the average spot intensity is divided by the average noise intensity. For SPR the average spot intensity is divided by the peak-noise sidelobe. The noise intensities are calculated with use of the entire 512×512-sample diffraction pattern region excluding a 128×128 window that just surrounds the spot array.

In previous studies we have used SPR and NU as our key measures of fidelity.[3–8] We also have reported SNR to provide comparisons with the work of other authors. However, since ED does not uniformly distribute noise of the full NRB, we also calculate a reduced-bandwidth SPR. For the modified SPR (SPR_m) the peak-noise intensity is found by using the intensity pattern that occupies the central 256×256 of the 512×512-sample image of the diffraction pattern.

In addition, the diffraction efficiency η is evaluated. However, rather than using the 512×512 FFT, we calculate the FFT of the 128×128 array directly without the zero padding. Efficiency is calculated as the sum of the intensities of 49 spots divided by the total energy in the 128×128-point diffraction pattern. This analysis does not take into account device-specific pixel aperture effects that determine the amount of energy diffracted into higher-order replicas that are outside the NRB of the diffraction pattern.

Gray-scale images of the intensity patterns are presented for several encodings to provide additional information on the generation of background noise. To bring out the background noise, we saturate the gray-scale level in each image so that full white corresponds to 3% of the average peak intensity. In each case, images of the entire NRB (i.e., all 512×512 samples) are shown.

4. COMPARISONS OF THE ENCODING ALGORITHMS

The results of encoding by ED, MDE, MD–PRE, and ED–PRE are presented in this section. Results for both on-axis and off-axis test functions are given. For each function the discussion first focuses on results derived with the single random sequence s_{ij} and preferred values of χ. Then the results are presented of the sensitivity analyses as a function of χ and for the ensemble of 21 random sequences.

For better appreciation of the property improvements possible with blended algorithms, we first present some reference designs using ED and MDE. Figure 3 shows how the background noise differs as a function of γ for the ED algorithm. For the ED algorithm as described in Ref. 17 the value of γ is presumably 1. The resulting diffraction pattern is shown in Fig. 3(a). The most pronounced noise appears at the corners of the 512 × 512-sample image. Faint noise peaks from the noise cloud extend out into the upper-left and lower-right corners of the (256 × 256-sample) reduced-bandwidth window. For $\gamma = 1.2$ the peak noise at the corners of the image [Fig. 3(b)] is reduced, and faint noise spikes appear over a larger extent of both the full NRB and the reduced-bandwidth window. An even more uniform distribution of noise spikes is seen in Fig. 3(c) for $\gamma = 1.3$. For γ near 1.6 the image [Fig. 3(d)] closely resembles the MDE design [Fig. 3(e)]. This result was unexpected. However, if γ is increased further, the diffraction pattern becomes quite distorted and no longer resembles the MDE design [Fig. 3(f)].

To gain some insight into the similarity between the results for MDE and ED with $\gamma = 1.6$, we produced histograms of the deviations $|\mathbf{a}_{c\,ij} - \mathbf{a}_{m\,ij}|$ between the encoded values produced by ED and by MDE, where the subscripts e and m designate values encoded by ED and MDE, respectively. Since the encoded values are on the unit circle, the largest deviation possible would be 2. For $\gamma = 1$ more than 50% of the deviations of the 16,384 pixels are larger than 0.6. For $\gamma = 1.1$ the histogram is nearly uniformly distributed over the full range from 0 to 2. However, as γ is increased further, the histograms have an increasing number of deviations near zero. For $\gamma = 1.5$ the deviations are less than 0.25 for 87% of the pixels, and there are deviations in excess of 0.6 for less than 4% of the pixels. For $\gamma = 1.6$ only 82% of the deviations are less than 0.25; but an even smaller amount, 1% of the deviations, exceed 0.6. The ED gray-scale image appears most similar to the MDE image when the number of large deviations are minimized. As γ is increased further, the percentage of large deviations increases to the point that the histogram becomes nearly uniformly distributed between 0 and 2. Thus for the encoding of the particular test function, the ED design tends to converge to, and produce a somewhat close approximation to, the MDE diffraction pattern for γ in the range 1.4–1.7.

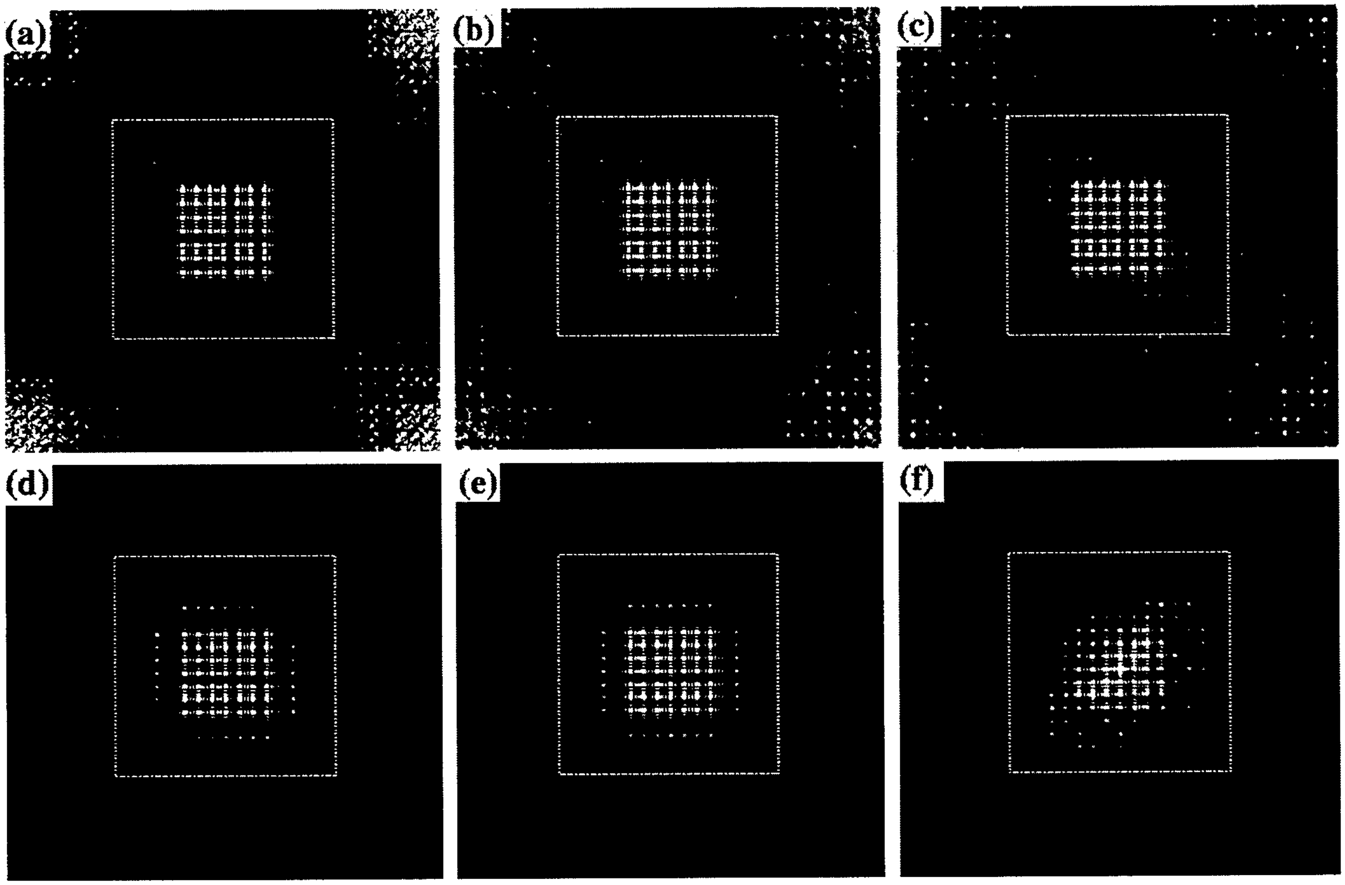


Fig. 3. On-axis diffraction patterns for encoding by (a)–(d) and (f) ED and (e) MDE. The value of γ used for ED is (a) 1.0, (b) 1.2, (c) 1.3, (d) 1.6, and (f) 5.5. The dotted square encloses the area used to calculate SPR_m. The gray-scale intensities are scaled so that full white corresponds to 3% of the average peak intensities of the 49 desired spots.

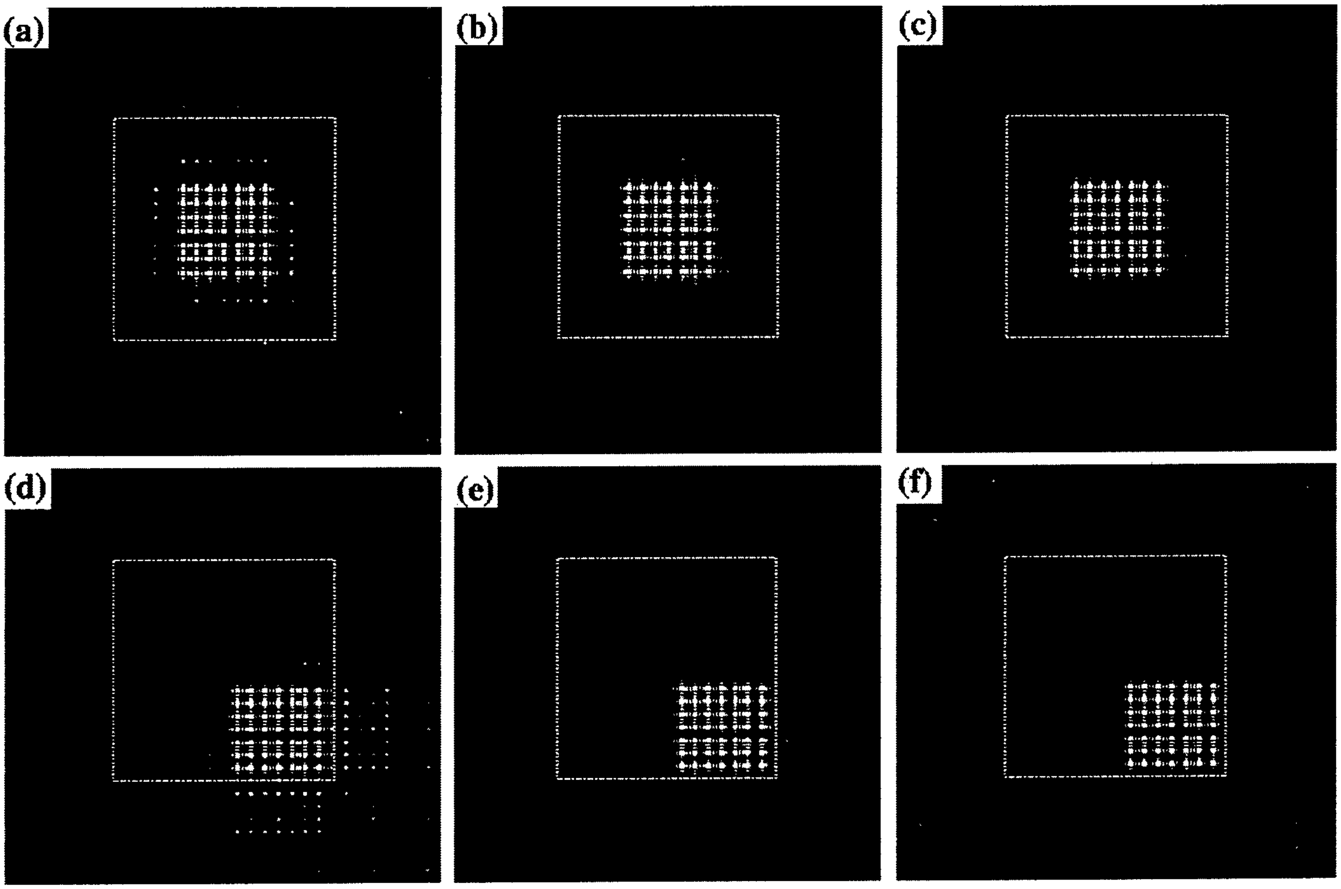


Fig. 4. (a)–(c) On-axis and (d)–(f) off-axis diffraction patterns for the maximum SPR design by (a) ED, $\gamma = 1.5$; (b) MD–PRE, $\gamma = 1.4$; (c) ED–PRE, $\gamma = 1.3$, $\chi = 0.6$; (d) ED, $\gamma = 1.5$; (e) MD–PRE, $\gamma = 1.4$; (f) ED–PRE, $\gamma = 1.3$, $\chi = 1$. The gray-scale normalization and the dotted square are identical to those used in Fig. 3.

A. Comparisons of On-Axis Designs

Figures 4(a)–4(c) show the background noise produced by ED, MD–PRE, and ED–PRE designs that have been tuned by γ and χ to produce the maximum value of SPR. Several performance metrics for each of these designs are reported in Table 1. For the ED algorithm with $\gamma = 1.5$ [Fig. 4(a)] the noise peaks appear to be even more uniformly distributed than for any of the designs in Fig. 3. For the MD–PRE algorithm [Fig. 4(b)] the noise background appears to be the most uniformly distributed of any result in Fig. 4. For the ED–PRE algorithm [Fig. 4(c)] the noise pattern demonstrates good features of both: of ED in that the noise is most pronounced in the corners of the image, but of MD–PRE in that the noise is more uniformly distributed over the corner regions than with ED. Thus ED–PRE appears to randomize the noise background, which reduces the noise peaks.

The background noise can be compared quantitatively in terms of SPR and SNR. Results for various encodings are presented in Fig. 5 and Table 1. Figure 5 presents both SPR and SPR_m (see Section 3) for ED, MD–PRE, and ED–PRE. For the ED design SPR_m is maximum for $\gamma = 1$. For MD–PRE at $\gamma = 1$ (i.e., the PRE algorithm) SPR_m is lower than for the ED design. However for $\gamma = 1.4$, SPR_m for MD–PRE is larger than for any ED design. The ED–PRE design results in an even larger value of this fidelity metric. For the full NRB the largest values of the SPR metric for ED–PRE and MD–PRE are essentially identical to those for SPR_m. In fact the entire SPR curve for MD–PRE is identical to the SPR_m curve. This is due to the presence of MDE-type sidelobes [e.g., in Fig. 3(b)] that are at the level of the random background noise. The SPR of the ED design is severely reduced (to a level even less than for MDE) by the inclusion of the noise peaks in the corner of the diffraction pattern.

Table 1. Best Encoding Performance for the On-Axis Function

Encoding Method	χ	γ	η(%)	SPR	SNR	NU(%)
Maximum SPR Design						
ED–PRE	0.6	1.3	72	53	784	3.3
ED	—	1.5	91	12	2660	10.2
MD–PRE	—	1.4	76	47	1000	5.8
Minimum NU Design						
ED–PRE	0.9	1.1	53	17	350	1.6
ED	—	1.2	58	4	421	5.2
MD–PRE	—	1.3	70	41	727	5.5
MDE[a]	—	∝	96	17	5220	19.1
PRE[b]	—	1.0	44	24	258	7.9

[a] For MDE γ is not an adjustable parameter.
[b] PRE has both best SPR and NU for γ equal to unity.

Therefore the blended algorithms provide a way to reduce peak noise over the full NRB by distributing the noise more uniformly.

The other key fidelity metric is NU. Table 1 shows that the ED–PRE design having the highest SPR also has a lower value of NU than does ED, MDE, or MD–PRE. Thus it outperforms the MD–PRE design in both SPR and NU despite having a somewhat lower SNR and diffraction efficiency η.

Table 1 also reports the performance of designs that produce the lowest overall value of NU. The ED–PRE algorithm produces an exceptionally low value of NU; however, SPR is now even lower than for MDE, and η is much lower than for the maximum SPR design. For the minimum-NU MD–PRE design γ is lower only by 0.1 than γ for the maximum SPR design, and all the performance differences between the two designs as measured by the metrics in Table 1 are only slight. The NU and SPR curves in Fig. 5 also suggest that MD–PRE is less sensitive to γ than is ED–PRE.

The results shown for ED–PRE in Fig. 5 are for the value of χ that produces either the highest value of SPR or the lowest value of NU. Figure 6 shows the range of variation of these performance metrics for values of $\chi \in [0.5, 1]$. The value of χ appears to have the greatest effect on NU, and the sensitivity decreases with increasing γ.

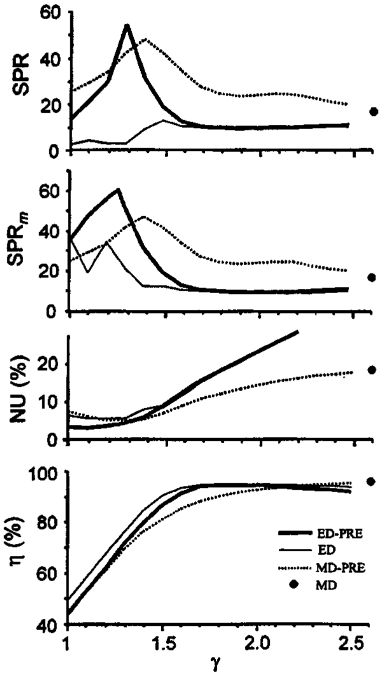


Fig. 5. Performance curves of the various encoding algorithms as a function of γ for the on-axis test function.[25] For ED–PRE the specific curve shown for NU is for $\chi = 0.9$, for SPR is for $\chi = 0.6$, and for SPR$_m$ is for $\chi = 0.7$. These curves achieve the best performance for ED–PRE as reported in Table 1.

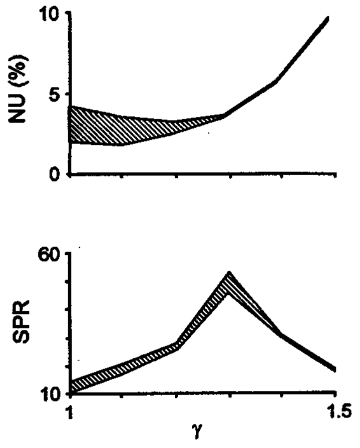


Fig. 6. Sensitivity of fidelity metrics of ED–PRE to the free parameter $\chi \in [0.5, 1]$ for the on-axis test function.

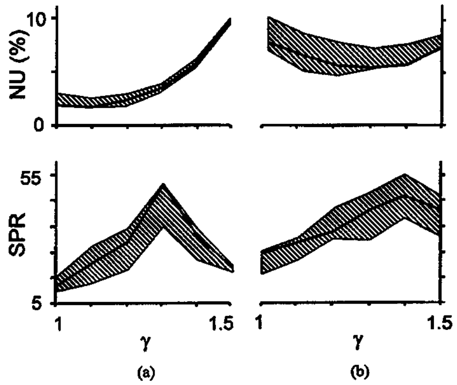


Fig. 7. Statistical variations of the fidelity metrics of (a) ED–PRE and (b) MD–PRE for an ensemble of encodings of the on-axis function. Shaded regions are bounded by the maximum and minimum values found for 21 random trials of each encoding algorithm. The respective curves from Fig. 5 are reproduced for comparison.

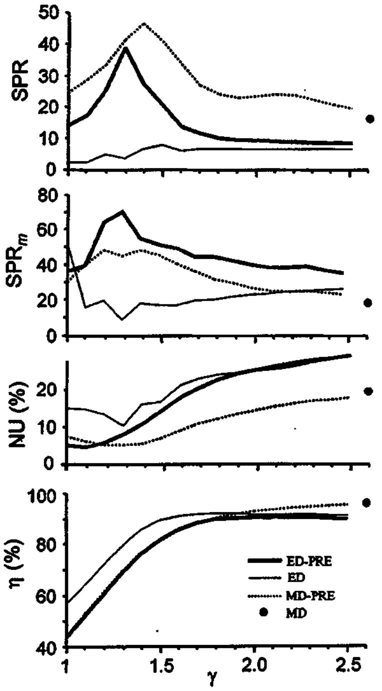


Fig. 8. Performance curves of the various encoding algorithms as a function of γ for the off-axis test function. For ED–PRE the specific curve shown for NU is for $\chi = 0.7$, for SPR is for $\chi = 1$, and for SPR$_m$ is for $\chi = 0.6$. These curves achieve the best performance for ED–PRE as reported in Table 2.

Figure 7 shows the range of variation of SPR and NU for an ensemble of the 21 random sequences of s_{ij}. The variation of ED–PRE is for the same values of χ as reported in Table 1. The ED–PRE curves from Fig. 5 are replotted for comparison. Whereas Fig. 5 and Table 1 report that ED–PRE produces higher SPR than MD–PRE, Fig. 7 shows for the ensemble that MD–PRE can produce a slightly higher value of SPR. A much larger ensemble is needed for us to be able to say whether MD–PRE outperforms ED–PRE in SPR on average and how frequently. Figure 7 also shows that ED–PRE can have substantially lower NU than MD–PRE and that NU for ED–PRE is much less sensitive to statistical variation than is MD–PRE.

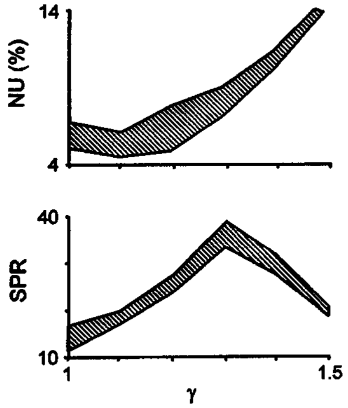


Fig. 9. Sensitivity of fidelity metrics of ED–PRE to the free parameter $\chi \in [0.5,1]$ for the off-axis test function.

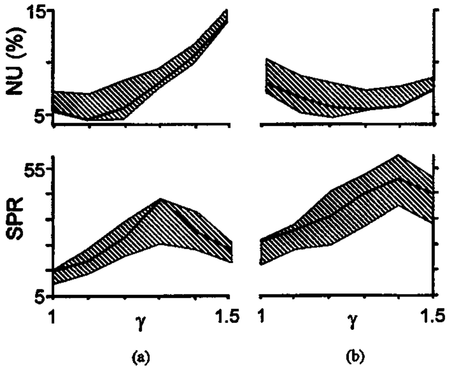


Fig. 10. Statistical variations of the fidelity metrics of (a) ED–PRE and (b) MD–PRE for an ensemble of encodings of the off-axis test function. Shaded regions are bounded by the maximum and minimum values found for 21 random trials of each encoding algorithm. The respective curves from Fig. 8 are reproduced for comparison.

Table 2. Best Encoding Performance for the Off-Axis Function

Encoding Method	χ	γ	η(%)	SPR	SNR	NU(%)
Maximum SPR Design						
ED–PRE	1.0	1.3	66	39	621	9
ED	—	1.5	59	8	2360	16.5
MD–PRE	—	1.4	76	47	1000	5.8
Minimum NU Design						
ED–PRE	0.7	1.1	52	18	340	4.5
ED	—	1.3	80	5	772	10.2
MD–PRE[a]	—	1.3	70	41	727	5.5

[a] MDE, PRE, and MD–PRE have identical performance for either off-axis or on-axis function. See Table 1 for MDE and PRE performance.

B. Comparisons of Off-Axis Designs

Figures 4(d)–4(f), Figs. 8–10, and Table 2 present the results for the off-axis design. These follow the same format as do Figs. 4(a)–4(c), Figs. 5–7, and Table 1. Figures 4(d), 4(e), and 4(f) presents the intensity diffraction patterns for ED, MD–PRE, and ED–PRE, respectively. The background noise pattern for each encoding method demonstrates textures (i.e., spiky for ED, white diffuse for MD–PRE, and colored diffuse for ED–PRE) similar to those shown in Figures 4(a)–4(c) for the on-axis design. Figure 8 and Table 2 compare the performance of the three algorithms. The major differences between the performance of the on-axis and the off-axis designs are that for the off-axis design (1) ED–PRE now clearly produces a much larger value maximum value of SPR_m than does MD–PRE; however, MD–PRE produces a much larger value of SPR than does ED–PRE; and (2) the minimum value of NU produced by ED–PRE is larger and is only slightly lower than the value of NU for MD–PRE. A major similarity is that the performance of off-axis and on-axis designs is nearly identical for MDE and MD–PRE. The only difference is in some values of SPR_m for MD–PRE. This is probably because in the off-axis design some of the MDE-type sidelobes lie outside the reduced-bandwidth region (which is the same area as was used for evaluating the on-axis designs).

These results point out one advantage of MD–PRE over ED–PRE: MD–PRE is less sensitive to where the desired pattern is centered. Certainly, other error diffusion kernels could be designed to center the noise reconstruction around the desired reconstruction. However this would further complicate the design process. Also, if the desired pattern is distributed over the full NRB, there may be no practical way to select ED weighting coefficients that spatially separate noise from the desired pattern. ED–PRE may provide little improvement in these cases.

For the sensitivity analyses (Fig. 9) it is found that NU and SPR for ED–PRE are both somewhat more sensitive to χ for the off-axis designs (Fig. 9) than for the on-axis designs (Fig. 6). In the statistical comparisons of ED–PRE with MD–PRE for the off-axis design, Fig. 10 shows that MD–PRE (for $\gamma = 1.4$) will almost always have a higher SPR than ED–PRE, while NU for MD–PRE is, at worst, only slightly higher than for ED–PRE. Slight differences between MD–PRE for on-axis and off-axis cases occurred because the ensemble of 21 random sequences used for calculating the curves in Fig. 7 was different from that used in Fig. 10.

5. CONCLUSIONS

We have introduced a new type of single-pixel encoding algorithm ED–PRE and have compared its performance with that of existing algorithms. The design criteria of principle concern, NU and SPR, emphasize fidelity rather than energy efficiency. Low SPR is sought for the entire NRB so that successive designs can address and utilize the entire SBWP available to the SLM. Depending on the specific test function encoded, the blended ED–PRE algorithm performs nearly as well as and sometimes better than the MD–PRE, and these blended algorithms substantially outperform the nonblended MDE, PRE, and ED algorithms. Even when the test function is centered and the performance metrics are calculated over a reduced-bandwidth window (which was the original intended application of the ED algorithm used here), the ED–PRE algorithm produces better performance than ED alone. Apparently ED–PRE uses the properties of ED to filter the background noise and distribute it nonuniformly over the diffraction plane, and it uses the properties of PRE to diffuse and reduce the peak intensity of the ED noise spikes.

In this paper we have also delineated the differences between the properties of ED and PRE. As a final distinction we note that both methods use diffusion, but it is used in different ways. In PRE the encoding error is scattered or diffused over the entire diffraction plane, forming a laser speckle pattern. In ED the encoding errors from one pixel are diffused forward into neighboring pixels in the modulation plane. The resulting noise background for ED is spiky rather than diffuse or speckled. One might choose to model the error sequence from ED as a stochastic process.[20] This is valid only to the extent that the sequence is described as a random process. However, for PRE the error sequence is a random process to the extent that a random-number generator represents a random process.

Although we have presented encoding algorithms specifically designed for continuous phase-only SLM's, it should be clear that ED–PRE can also be developed both for quantized and for coupled modulation characteristics by blending the approaches presented in Refs. 5, 8, 17. We would expect the background noise to have an appearance and the performance metrics to show dependencies on the free parameters γ and χ similar to those observed for phase-only modulation. However, it would be interesting to find out whether the relative performance differences between ED, ED–PRE, and MD–PRE are maintained for different modulation characteristics. We have no reason to suspect that this either might or might not be the case.

In summary, ED–PRE can improve on the fidelity of ED alone. We suspect that ED–PRE (as illustrated by the two examples presented here) is competitive with MD–PRE and that in some cases it may even outperform MD–PRE.

ACKNOWLEDGMENTS

This study was supported by U.S. Office of Naval Research contract N00014-96-1-1296 and NASA cooperative agreement NCC5-222.

*When the work was performed the authors were with University of Louisville. M. Duelli's current address is Optical Coating Laboratory, Inc., 2789 Northpoint Parkway, MS 125-3, Santa Rosa, California 95407-7397.

REFERENCES AND NOTES

1. R. W. Cohn and M. Liang, "Approximating fully complex spatial modulation with pseudorandom phase-only modulation," Appl. Opt. **33**, 4406–4415 (1994).
2. L. G. Hassebrook, M. E. Lhamon, R. C. Daley, R. W. Cohn, and M. Liang, "Random phase encoding of composite fully complex filters," Opt. Lett. **21**, 272–274 (1996).
3. R. W. Cohn and W. Liu, "Pseudorandom encoding of fully complex modulation to bi-amplitude phase modulators," in *Diffractive Optics and Micro-Optics*, Vol. 5 of 1996 OSA Technical Digest Series (Optical Society of America, Washington, D.C., 1996), pp. 237–240.
4. R. W. Cohn and M. Liang, "Pseudorandom phase-only encoding of real-time spatial light modulators," Appl. Opt. **35**, 2488–2497 (1996).
5. R. W. Cohn, "Pseudorandom encoding of complex valued functions onto amplitude coupled phase modulators," J. Opt. Soc. Am. A **15**, 868–883 (1998).
6. R. W. Cohn and M. Duelli, "Ternary pseudorandom encoding of Fourier transform holograms," J. Opt. Soc. Am. A **16**, 71–84 (1999); Errata, 1089–1090 (1999).
7. R. W. Cohn, "Analyzing the encoding range of amplitude-phase coupled spatial light modulators," Opt. Eng. **38**, 361–367 (1999).
8. M. Duelli, M. Reece, and R. W. Cohn, "A modified minimum-distance criterion for blended random and nonrandom encoding," J. Opt. Soc. Am. A **16**, 2425–2438 (1999).
9. R. W. Cohn and L. G. Hassebrook, "Representations of fully complex functions on real-time spatial light modulators," in *Optical Information Processing*, F. T. S. Yu and S. Jutamulia, eds. (Cambridge U. Press, Cambridge, UK, 1998), Chap. 15, pp. 396–432.
10. B. R. Brown and A. W. Lohmann, "Complex spatial filter," Appl. Opt. **5**, 967–969 (1966).
11. W. J. Dallas, "Computer-generated holograms," in *The Computer in Optical Research*, B. R. Frieden, ed. (Springer-Verlag, Berlin, 1980), Chap. 6, pp. 291–366.
12. R. D. Juday, "Optimal realizable filters and the minimum Euclidean distance principle," Appl. Opt. **32**, 5100–5111 (1993).
13. L. B. Lesem, P. M. Hirsch, and J. A. Jordon, Jr., "The kinoform: a new wavefront reconstruction device," IBM J. Res. Dev. **13**, 150–155 (1969).
14. J. L. Horner and P. D. Gianino, "Phase-only matched filtering," Appl. Opt. **23**, 812–816 (1984).
15. R. W. Floyd and L. Steinberg, "An adaptive algorithm for spatial grayscale," Proc. Soc. Inf. Disp. **17**, 78–84 (1976).
16. R. Hauk and O. Bryngdahl, "Computer-generated holograms with pulse density modulation," J. Opt. Soc. Am. A **1**, 5–10 (1984).
17. S. Weissbach, F. Wyrowski, and O. Bryngdahl, "Digital phase holograms: coding and quantization with an error diffusion concept," Opt. Commun. **72**, 37–41 (1989).
18. S. Weissbach and F. Wyrowski, "Error diffusion procedure: theory and applications in optical signal processing," Appl. Opt. **31**, 2518–2534 (1992).
19. A. G. Kirk, A. K. Powell, and T. J. Hall, "The design of quasi-periodic Fourier plane array generators," in *Optical Information Technology*, S. D. Smith and R. F. Neale, eds. (Springer-Verlag, Berlin, 1993), pp. 47–56.
20. E. Barnard, "Optimal error diffusion for computer-generated holograms," J. Opt. Soc. Am. A **5**, 1803–1811 (1988).
21. M. Duelli and R. W. Cohn, "Pseudorandom encoding for real-valued ternary spatial light modulators," Appl. Opt. **38**, 3804–3809 (1999).
22. F. Wyrowski, "Upper bound of the diffraction efficiency of

diffractive phase elements," Opt. Lett. **16**, 1915–1917 (1991).

23. J. A. Davis and D. M. Cottrell, "Random mask encoding of multiplexed phase-only and binary phase-only filters," Opt. Lett. **19**, 496–498 (1994).
24. J. M. Goodman, "Effects of film nonlinearities," in *Introduction to Fourier Optics* (McGraw-Hill, San Francisco, Calif., 1968), Sec. 8-6, pp. 230–241.
25. To maintain as much consistency as possible in comparing all the curves and tables and to avoid excessive computation, we have calculated and reported all performance metrics for values of $\gamma = 1, 1.1, 1.2, \ldots$. This results in adequately smooth and sampled curves except in one case. For the SPR_m curve of ED–PRE in Fig. 5, finer sampling led to a significant increase in SPR_m, from 53 at $\gamma = 1.2$ and 1.3 to 60 at $\gamma = 1.26$. This additional point is included in the plot in Fig. 5. We also checked the maxima of other SPR and SPR_m performance curves, using finer sampling increments. However, since the change in appearance is minimal and the maximum values of the curves would change by no more than a few tenths, these additional findings are omitted.

Nonlinear effects of phase blurring on Fourier transform holograms

Markus Duelli,* Li Ge, and Robert W. Cohn

The ElectroOptics Research Institute, University of Louisville, Louisville, Kentucky 40292

Received October 22, 1999; revised manuscript received May 1, 2000; accepted May 8, 2000

Liquid-crystal light valves can have intensity-dependent resolution. We find for a nematic liquid-crystal light valve that this effect is well modeled as a phase that has been blurred by a linear space-invariant filter. The phase point-spread function is measured and is used in simulations to demonstrate that it introduces intermodulation products to the diffraction patterns of computer-generated Fourier transform holograms. Also, the influence of phase blurring on a pseudorandom-encoding algorithm is evaluated in closed form. This analysis applied to a spot array generator design indicates that nonlinear effects are negligible only if the diameter of the point-spread function is a small fraction of the pixel spacing. © 2000 Optical Society of America [S0740-3232(00)01608-2]

OCIS codes: 230.6120, 070.2580, 030.6600, 160.3710, 100.3020.

1. INTRODUCTION

Loss of spatial resolution in a linear space-invariant imaging system is determined by the convolution of the input image with the point spread function (PSF) of the system. That is, the PSF blurs the input image. Many spatial light modulators (SLM) also can be viewed as linear space-invariant systems that convert an input image into an output image. Image blurring from this transformation can also be ascribed to a device PSF (sometimes referred to as an influence function).

For liquid-crystal light valves (LCLV), resolution has been reported to depend on the input intensity.[1–4] For instance, we recently studied a LCLV that was quoted as having 40 line pairs/mm resolution for low illumination levels and 4 line pairs/mm for high illumination levels.[5] Thus a simple convolution model of resolution loss is not appropriate for LCLV's used as intensity displays. However, a convolution relationship appears to exist between input intensity images and output phase images. Specifically, in this paper we measure the phase PSF and find that, except for a scale factor, it is independent of the input intensity. Therefore phase blurring in LCLV's can be modeled by linear space-invariant filtering of the phase. The reason for the apparent loss in output resolution as a function of input intensity in Refs. 1–5 is not due to a loss in input phase (ϕ) resolution but rather to the nonlinear transformation from phase to the resulting complex-valued modulation $[\exp(j\phi)]$.

In this study we specifically consider the nonlinear effects of phase blurring on phase-only LCLV's used to produce optical Fourier transforms. It is important to recognize that even if there is no phase blurring, the nonlinear transform $\phi \rightarrow \exp(j\phi)$ is inherent in the design of phase-only computer-generated holograms. Various encoding techniques have been developed for which the Fourier transform of $\exp(j\phi)$ approximates the Fourier transform of a desired fully complex modulation $\mathbf{a}_c \equiv a_c \exp(j\psi)$.[6,7] One such encoding method, pseudorandom encoding,[5,8–11] (PRE), approximates the desired modulation in an average sense (which is reviewed in Subsection 4.A). Here we note that this approximate mapping from the desired signal $\mathbf{a}_c$ to the encoded phase-only signal $\exp(j\phi)$ can be viewed as a linear (or more precisely, a quasi-linear) space-invariant system. In this way encoding approximately linearizes a nonlinear system.

This quasi-linear relationship found for encoding algorithms can be destroyed by phase blurring. The simplest model necessary to show that phase blurring is a nonlinear effect is to convolve the phase $\phi(x)$ that is a function of spatial coordinate x with the single-lag filter function (or phase PSF)

$$h(x) = (1 - \alpha)\delta(x) + \alpha\delta(x - \Delta), \tag{1}$$

where $\delta(x)$ is the Dirac delta function, Δ is a spatial offset, and α is a weighting coefficient between 0 and 1. The frequency response of the blurred phase is the frequency response of the phase multiplied by the frequency response of the PSF. However, the complex-valued modulation becomes

$$\exp[j\phi(x)*h(x)] = \exp[j\phi(x)] \times \exp\{j\alpha[\phi(x - \Delta) - \phi(x)]\}, \tag{2}$$

and its Fourier transform contains the encoded spectrum of $\exp(j\phi)$ convolved, rather than multiplied, by the spectrum of $\exp\{j\alpha[\phi(x - \Delta)-\phi(x)]\}$. This additional term is responsible for errors in the intensities of the desired Fraunhofer diffraction pattern at the design frequencies and for noticeable unwanted diffraction orders at the other frequencies. Therefore phase blurring introduces nonlinear effects into the complex-valued modulation and the resulting Fraunhofer diffraction pattern. In Sections 4 and 5 we will use the single-lag blur model of Eq. (1) in modeling and experimental demonstrations of the blurring effect. This model is especially well suited for ex-

0740-3232/2000/091594-12$15.00

perimental implementation of blurring with a pixelated, electrically addressed SLM that is reported in Section 5.

Phase blurring also bears some resemblance to the problem of phase scaling errors in computer-generated holograms (CGH's). This can be seen by rewriting Eq. (2) as

$$\exp[j\phi(x)*h(x)] = \exp[j(1-\alpha)\phi(x)]\exp[j\alpha\phi(x-\Delta)]. \tag{3}$$

The term $\exp[j(1-\alpha)\phi(x)]$ has scaled phase, which is known to reduce diffraction efficiency and for off-axis holograms introduces an on-axis component.[7,12,13]

Unwanted diffraction orders arising from a phase-blurred LCLV have been experimentally observed for both PRE and a nonrandom phase-only encoding algorithm.[9] In Ref. 9 and earlier studies, while we were aware of a spatial-frequency-dependent loss of phase range, we did not model this as phase blurring, nor did we consider the related nonlinear effects on the diffraction pattern.[5,8,9] Instead, we attempted to minimize phase blurring by making the pixel spacing of the discretely sampled CGH's large with respect to the maximum phase slope anticipated. At the time, the major source of errors between the experimentally measured and the simulated results was assumed to be various point nonlinearities—e.g., the inaccurate setting of the mapping between input intensity and phase, variations in response across the SLM, and quantized phase levels. However, in this study we find that even a small amount of blurring can be quite noticeable. For the spot array generator designs considered in this paper, we find that the diameter of the phase PSF needs to be a quite small fraction of the pixel spacing for the nonlinear effects of blurring to be negligible.

The phase-blurring paradigm provides a unified view of SLM spatial properties that have been characterized in terms of spatial-frequency-dependent phase (Fig. 4 of Ref. 8), diffraction efficiency (Refs. 1, 2, 4, 14, and Fig. 5 of Ref. 8), and effective complex amplitude [Fig. 11(b) of Ref. 5]. Phase-blurring models have an added advantage in that they can be used to model distortion in phase-only holograms that consist of a multitude of spatial frequencies rather than a single spatial frequency.

In this paper we demonstrate through a combination of measurements and simulations that phase-only LCLV's are reasonably well described by a space-invariant phase-blurring model, and we quantify the magnitude of the effect on Fourier transform holograms designed by PRE. We also introduce blurring into an electrically addressed SLM and evaluate the performance as a function of the degree of blurring. Although phase blurring affects all encoding algorithms, we focus on PRE because the effects of phase blurring on PRE can be analyzed in closed form. Such an expression is derived for the case of single-lag blurring and is used to evaluate the distortion of the far-field pattern from a PRE designed spot array generator. The expression is also used to consider the possibility of predistorting the phase so as to compensate for blurring.

2. ILLUSTRATION OF THE EFFECTS OF PHASE BLURRING

The effects of phase blurring on Fourier transform holograms are illustrated in Fig. 1 for the comparison of a spot array generator that does not suffer from phase blurring [Fig. 1(a)] with one that does [Fig. 1(b)]. For diffractive optical elements (DOE's) that have abrupt transitions between pixels, blurring should not be an issue. For example, Fig. 1(a) shows the far-field intensity pattern (for an 850-nm illumination source) of an eight-phase-level transmissive DOE that was designed by a specific blended PRE algorithm.[10,15] The device consists of 300 × 300 square pixels. Each pixel is 13.3 μm on a side, and the sidewalls are essentially vertical except for some ledges that are due to misalignment errors between successive mask layers. (The ledges, as measured with an atomic force microscope are never larger than 0.4 μm). The diffractive optic is designed to produce an 8×8 array of equally spaced spots off axis. One unit spacing to the right and one unit below the spot array is a faint spot located on the optical axis. Additional features are the sidelobes between the spots. These are due to the small number of repetitions (4×4) of the 75×75 pixel unit cell in the spot array design function. Finally, there is a background pattern of speckle that is an essential byproduct of the PRE algorithm. There are no other noticeable

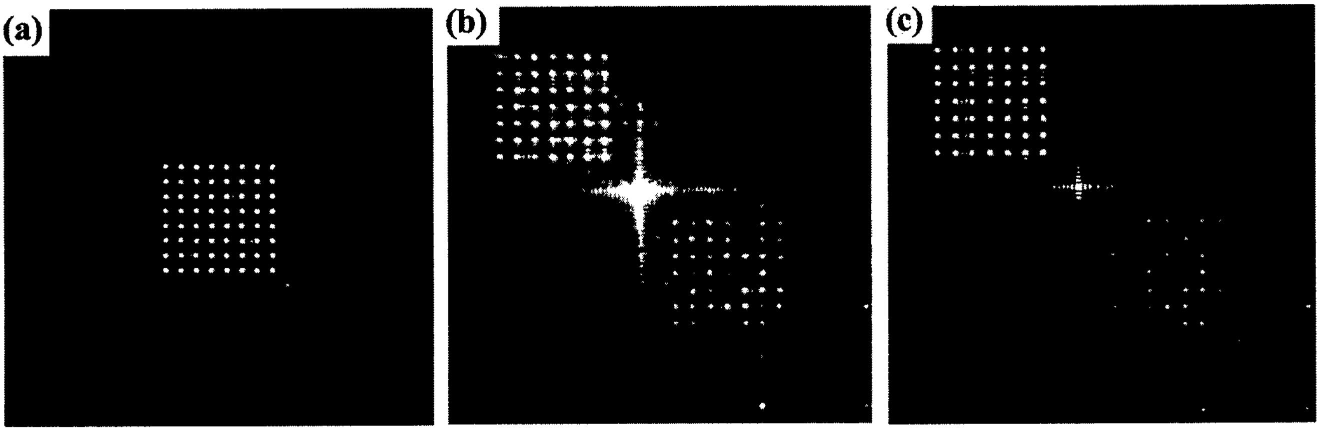


Fig. 1. Far-field diffraction patterns from (a) a DOE, (b) a LCLV, and (c) a simulation of (b) that includes a linear shift-invariant phase-blurring model. Maximum white in the gray-scale images correspond in (a) and (c) to 30% and in (b) to 20% of the average of the peak intensities of the spots in the desired spot array.

Table 1. Performance of PRE Implemented on a DOE

Metric	Theory	Experiment
SNR	916	942
SPR	35	52
NU[a] (%)	8.2	8.3
η (%)	38	41

[a] The experimental result for NU is the average for seven devices measured with the procedure described in Ref. 16. The standard deviation of NU is 0.8%.

Table 2. Performance of PRE Implemented on a LCLV

Metric	Ideal (No Blurring)	Simulated Phase Blurring	Experimental Results	Prefiltering and Blurring
SNR	254	148	73	340
SPR	17	3	1.9	18
NU (%)	10	27	33	18
η (%)	43	31	—	49
E_0/E_1 (%)	0	59	—	2
E_{-1}/E_1 (%)	0	15	36	0

features between the optical axis and the (1,1) grating order (corresponding to the spatial frequency that is the reciprocal of the pixel pitch).

Figure 1(b) shows a 7 × 7 spot array from a Hughes LCLV (nematic, parallel aligned). The modulation pattern consists of a 128 × 128 pixel image designed by PRE. The modulation values are identical to those reported in Ref. 9. The modulation is generated by projecting a gray-scale image from a red phosphor CRT onto the write side of the LCLV. The CRT is driven by the signal from a computer video display card set to a resolution of 800 × 600 pixels and a frequency of 56 Hz. A subimage of 384 × 384 video pixels is imaged into a 19.2-mm × 19.2-mm area of the LCLV. Each modulation pixel corresponds to 3 × 3 video pixels or 150 μm × 150 μm. The modulation pattern is read out in phase-only mode by reflecting a linearly polarized 488-nm laser beam off the read side of the SLM. Additional information on the optical setup and the LCLV characteristics are described in Refs. 5 and 9.

Many of the same features as in Fig. 1(a) are seen in Fig. 1(b). Note that the pattern in Fig. 1(b) has a higher level of background speckle than that in Fig. 1(a), but this is due to the smaller number of pixels in the PRE design. The main differences between the two patterns are that in Fig. 1(b) there are a relatively intense on-axis spot, a faint mirror image, and an additional pattern of spots to the right and below the mirror image. If the spot array is designed centered on axis, these patterns are coincident, and if the spot array is further off axis, the patterns can be separated into nonoverlapping diffraction orders (which is discussed further in Subsection 5.D).

The simulated performance of the 8 × 8 spot array in Table 1 and of the 7 × 7 spot array in the first column of Table 2 are somewhat comparable in terms of diffraction efficiency (η), signal-to-noise ratio (SNR), signal-to-peak-noise ratio (SPR) and nonuniformity (NU). (See Appendix A for a review and definition of these metrics.) In Table 1 the experimentally measured performance of the glass diffractive optic is comparable to the simulated performance. However, in Table 2 the measured performance is substantially different and degraded for the design that is implemented on the LCLV. The performance change appears to be due in large part to the phase PSF of the LCLV, as is considered in the next section.

3. EVALUATION OF PHASE BLURRING AND ITS INFLUENCE ON PERFORMANCE

We evaluated the effect of the phase PSF by convolving the experimentally measured PSF with the desired phase. We first measure the PSF by focusing a He–Ne laser beam onto the write side of the LCLV device using a 6× microscope objective. The waist diameter of the Gaussian beam is 11.1 μm full width at half-maximum (FWHM, or 32 μm at the e^{-2} intensity level). The LCLV pattern is read out as in Section 2 with a 488-nm wavelength. The beam is interfered with a reference wave front in a Michelson interferometer. The interferogram is recorded on a CCD camera for various write beam powers between 1.3 and 6.8 μW, corresponding to peak phase shifts between 0.3π and 2π. The phase profile for each image is calculated by point-by-point conversion from the intensity to phase. For each resulting image the phase profile is approximately a circular Gaussian with a diameter of 54 μm FWHM. (For comparison, the interferogram intensity pattern has a diameter of 72 μm FWHM when the peak phase shift is π.) The unchanging shape of the phase profile indicates that the phase profile corresponds to the phase PSF and that phase blurring for the LCLV is reasonably modeled by linear space-invariant filtering.

We also verified that the divergence of the write beam through the photodetecting layer of the LCLV does not significantly contribute to blurring. This was demonstrated by observing that the interferogram remains unchanged when the waist position is translated several millimeters along the optical axis. [Note that the theoretical depth of focus (range over which the waist expands less than a factor of 1.41×) is 2.54 mm.] These results show that the width of the phase PSF is nearly that of the 150-μm modulation pixels used for the experiments in Section 2.

While these experiments demonstrate the presence of phase blurring of the LCLV, they do not include the additional sources of resolution loss from the video card, the CRT, and the imaging system that are part of the complete SLM system described in Section 2. For this reason the phase PSF also is measured for the entire system with a single video pixel used as our closest approximation to a point source. The geometric image of this pixel (based on a 1.9× demagnification between the CRT and the LCLV) would be 50 μm × 50 μm. The resulting phase image is once again observed to be approximately circular Gaussian but with a somewhat larger diameter of 59.4 μm FWHM. In our computer simulations we use this phase pattern as an approximation of the phase PSF of the SLM system.

The effect of blurring is evaluated by convolving the measured PSF with the desired phase modulation and then Fourier transforming the blurred modulation. This is done by digital simulation in which the PSF and the designed modulation are sampled every 37.5 μm in the x and y directions corresponding to a sample spacing that is one fourth the modulation pixel spacing $\Delta = 150\ \mu$m. A 7×7 array of samples is used to represent the phase PSF. Sample values outside the 7×7 array are rather small and are treated as zero in the simulations. Note that the sample values are normalized so that they add up to unity and thus reproduce the desired phase when convolved with a constant phase image. The complex modulation is calculated from the resulting blurred phase. This is zero padded to produce 2048×2048 sample points, and then the simulated diffraction pattern is calculated with the fast Fourier transform (FFT).

This procedure was used to model the experimental results that are shown in Fig. 1(b). Using the same designed modulation in the simulation procedure gives the diffraction pattern of Fig. 1(c). The location and relative strengths of the unwanted diffraction orders appear to the eye to be quite similar. The major disagreement is that the actual on-axis spot is much brighter than the simulated spot. This is due to reflections from the cover glass of the LCLV, which is not incorporated into the models. The second and third columns of Table 2 compare three performance measures of the simulated and measured spot arrays. For each measure the experimental performance is lower than for the simulated performance. However, these values compare much more closely than they do with the values for the spot array that is unaffected by blurring (first column of Table 2). We have repeated these comparisons for a number of designs and encoding algorithms, and we observe similar trends in each case. Thus we believe that these results indicate that blurring is a major contributor to the loss of performance.

To gain additional information on the effect of phase blurring, we repeated the simulations with PSF's of various diameters between 6 and 90 μm. The same measured PSF as above is used except that it is resampled and scaled to the corresponding diameter. The sample spacing (for both the PSF and modulation) is also reduced from one fourth to one eighth of the modulation pixel spacing Δ to permit adequate sampling of the smaller-diameter PSF's. For reasons of numerical efficiency the PSF kernel is limited to an 11×11 array of samples. For scalings of the PSF to large diameters the truncation of the tails of the PSF will lead to an underestimation of the effects of blurring, and for very large diameters the curves eventually flatten out as a result of the truncated PSF approaching a rect function. Even for $\Delta_{\mathrm{PSF}}/\Delta = 0.396$ the spatial extent of the kernel is somewhat shorter than the kernel used for the simulated results in Table 2. This leads to NU and E_0/E_1 being somewhat smaller in Fig. 2 than in Table 2. However, the discussion of the Fig. 2 results will focus on the smaller values of $\Delta_{\mathrm{PSF}}/\Delta$ for which the truncation effect is even less significant.

Figure 2 summarizes the results of these simulations for two performance measures: The relative energy in the on-axis spot relative to the resulting spot array (E_0/E_1) and NU of the spot array. These are plotted as a function of PSF diameter relative to pixel spacing ($\Delta_{\mathrm{PSF}}/\Delta$). The top plot in Fig. 2 shows that an increasing fraction of the energy appears in the on-axis spot with increased blurring. For a relative PSF diameter of only 15%, the energy in the on-axis spot is 10% of the energy in the spot array, or, equivalently, ~5× brighter than the average spot in the 49-spot array. The bottom plot in Fig. 2 shows that the spot array intensities become less uniform with increasing blur diameter. For a relative PSF diameter of 15%, NU has increased by 20% relative to NU for the ideal design. Even for this small degree of blurring, the changes to this design are quite significant. From these results we conclude that the spot array generator is surprisingly sensitive to a relatively small amount of blurring.

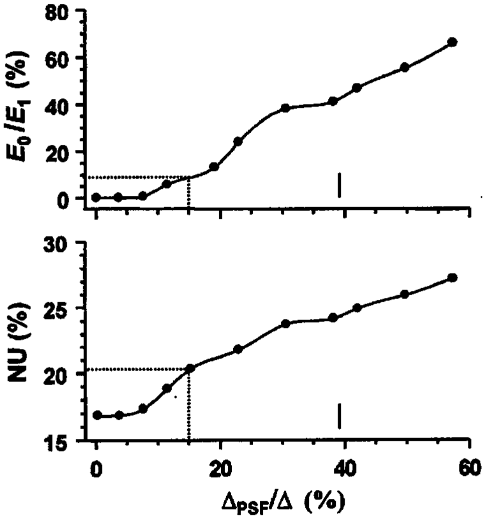


Fig. 2. Simulated effect of phase blurring on performance for phase PSF's of various diameters. The vertical bars indicate the value of $\Delta_{\mathrm{PSF}}/\Delta$ corresponding to the phase PSF measured for the actual LCLV. The dotted lines indicate the performance values for $\Delta_{\mathrm{PSF}}/\Delta = 0.15$.

4. ANALYSIS OF PHASE BLURRING ON PSEUDORANDOM ENCODING

The results of Section 3 indicate that phase blurring can introduce undesirable and noticeable nonlinear effects. PRE algorithms,[8,9,11] which approximate linear mappings between the desired complex modulation values and the modulation values that the SLM can actually produce, are subject to these nonlinear effects. The influence of a two-pixel (nearest-neighbor) blurring model on an encoding algorithm is analyzed in this section. In Section 5 this class of phase PSF's is implemented and experimentally studied with an electrically addressed SLM for which blurring is negligible. This permits a comparison of measured and simulated diffraction patterns for various degrees of blurring.

A. Derivation and Analysis of the Pseudorandom-Encoding Algorithm

The derivation of the blurring-induced distortion follows from the definitions and properties of PRE, which we review. PRE algorithms for modulation-range-limited SLM's encode the desired complex modulation $\mathbf{a}_{ci}$ at the ith SLM pixel in the average sense,

$$\mathbf{a}_{ci} = \int \mathbf{a} p_i(\mathbf{a})/d\mathbf{a} \equiv \langle \mathbf{a} \rangle_i, \qquad (4)$$

where $p_i(\mathbf{a})$ is the probability density function (PDF) of the SLM modulation $\mathbf{a}$ that is a random variable and $\langle \cdot \rangle$ is defined as the ensemble-average operator. The encoding algorithm is designed by finding a PDF that satisfies the integral equation Eq. (4). For phase-only SLM's for which $\mathbf{a}_i = \exp(j\phi_i)$, Eq. (4) simplifies to

$$\mathbf{a}_{ci} = \int \exp(j\phi) p_i(\phi) \mathrm{d}\phi = \langle \exp(\phi) \rangle_i. \qquad (5)$$

This can be satisfied by a variety of PDF's as long as the magnitude of the desired complex modulation $\mathbf{a}_{ci}$ is less than unity.[16]

A PRE algorithm that is amenable to analysis of phase blurring is based on the binary-phase random variable

$$\phi_i = \begin{cases} \psi_i - \nu_i/2, & 0 \leq s_i < 1/2 \\ \psi_i + \nu_i/2, & 1/2 < s_i \leq 1' \end{cases} \qquad (6)$$

where $\psi_i = \arg(\mathbf{a}_{ci})$ is the phase of the desired complex modulation, $\nu_i/2$ is a random binary-phase offset, and s_i is a uniform random variable between 0 and 1. The PDF for the phase random variable in Eq. (6) is then written as

$$p_i(\phi) = \tfrac{1}{2}\{\delta[\phi - (\psi_i - \nu_i/2)] + \delta[\phi - (\psi_i + \nu_i/2)]\}, \qquad (7)$$

where $\delta(\phi)$ is the Dirac delta function. Evaluating the expected complex value of Eq. (5) with Eq. (7) gives

$$\langle \mathbf{a} \rangle_i = \cos(\nu_i/2)\exp(j\psi_i). \qquad (8)$$

If the value of phase offset is set to

$$\nu_i/2 = \arccos(|\mathbf{a}_{ci}|), \qquad (9)$$

then the result sought for Eq. (4), $\mathbf{a}_{ci} = \langle \mathbf{a} \rangle_i$, is obtained. This result leads to an encoding formula in which the magnitude of the random phase offset is set by using Eq. (9); and then, with use of Eq. (6), the sign of the phase offset is randomly selected according to the value of the random number s_i. The offset is added to the desired encoded phase $\psi_i = \arg(\mathbf{a}_{ci})$ to produce the encoded phase ϕ_i. This process is repeated for each pixel of the N pixels of the SLM.

The resulting far-field diffraction pattern of the complex field is proportional to

$$\mathbf{A}(f_x) \equiv \sum_{i=1}^{N} \mathbf{a}_i \exp(-j2\pi i \Delta f_x), \qquad (10)$$

which is the Fourier transform of an array of equally spaced point sources of pitch Δ. (For purposes of explanation, the modulation is described in only one dimension and the pixel apertures are considered to be infinitesimal in width.) The expected value of Eq. (10) gives the desired complex diffraction pattern,

$$\mathbf{A}_c(f_x) = \langle \mathbf{A}(f_x) \rangle = \sum_{i=1}^{N} \mathbf{a}_{ci} \exp(-j2\pi i \Delta f_x), \qquad (11)$$

where $\mathbf{A}_{ci}(f_x)$ is the Fourier transform of the desired modulation. The far-field diffraction pattern of the encoded modulation is known to produce a noise-perturbed approximation to the diffraction pattern that would result from the desired complex-valued modulation $\mathbf{a}_{ci}$.[11] The presence of background noise is indicated in the expected intensity pattern:

$$\begin{aligned}\langle |\mathbf{A}(f_x)|^2 \rangle &= \sum_{i=1}^{N} \sum_{k=1}^{N} \langle \mathbf{a}_i \mathbf{a}_k^* \rangle \exp[-j2\pi(i - k)\Delta f_x] \\ &= |\mathbf{A}_c(f_x)|^2 + \sum_{i=1}^{N} (1 - |\mathbf{a}_{ci}|^2). \end{aligned} \qquad (12)$$

The second equality identifies the desired power spectrum and an additional white background noise. This result follows for the specific condition that $\mathbf{a}_i$ is statistically independent of $\mathbf{a}_k$ for $i \neq k$. (Independence is imposed in the design of PRE algorithms to simplify their derivation and implementation. In fact, blurring introduces statistical dependence between the values of neighboring pixels, which complicates the expected intensity, as will be shown in Subsection 4.B.) In the first equality, if the terms $\langle |\mathbf{a}_i|^2 \rangle = 1$ were replaced with $|\langle \mathbf{a}_i \rangle^2 \equiv |\mathbf{a}_{ci}|^2$, then Eq. (12) would equal $|\mathbf{A}_c(f_x)|^2$. This factorization has been performed to produce the second equality.

The single summation term in Eq. (12) corresponds to the average intensity level of white background noise in the diffraction pattern. The noise is observed in experiments and simulations to be a speckle pattern.[8,11] The intensity of the noise pattern depends on the intensity of the desired modulation values $\mathbf{a}_{ci}$. The closer the values are to unity magnitude, the lower is the intensity of the noise pattern. This can be viewed as measuring the dissimilarity between the desired modulation and the modulation achievable with the particular SLM. For this reason we often refer to this term as the error signal.

The expectation of the squared intensity provides additional information: specifically, the statistical variations of the pattern. The expression is derived in Ref. 11. Analyses of this higher-order moment show that the magnitude of the error signal is closely related to the deviations between the desired and the resulting diffraction patterns.

B. Effect of Blurring on the Encoding Algorithm

The effect of blurring on the PRE algorithm of Subsection 4.A is derived under the assumption that the encoded phase ϕ_i is blurred by a discrete version of the blurring function of Eq. (1). The two-pixel influence function is chosen because (1) it is the simplest blurring function to evaluate theoretically, (2) it can be directly implemented with an available electrically addressed SLM, and (3) it introduces the basic nonlinear effects that would be produced by a more extended blurring function. Repeating the analysis of Subsection 4.A for a two-pixel blur function evaluates to an effective complex amplitude of

$$\begin{aligned}\mathbf{b}_i &\equiv \langle \exp\{j[(1-\alpha)\phi_i + \alpha\phi_{i-1}]\}\rangle \\ &= \langle \exp[j(1-\alpha)\phi_i]\rangle\langle \exp(j\alpha\phi_{i-1})\rangle \\ &= \cos[(1-\alpha)\nu_i/2]\cos(\alpha\nu_{i-1}/2) \\ &\quad\times \exp\{j[(1-\alpha)\psi_i + \alpha\psi_{i-1}]\}. \qquad (13)\end{aligned}$$

The second equality follows from the statistical independence of the random variables ϕ_i. Equation (13) reflects the distortion introduced into Eq. (8) by blurring. The desired phase ψ_i from Eq. (8) is filtered by Eq. (1) in Eq. (13). Also, the amplitude is distorted. The Fourier transform [see Eq. (11)] of the sequence in Eq. (13) produces $\mathbf{B}(f_x)$, which is the expected complex diffraction pattern. The expected intensity diffraction pattern due to blurring is derived following a factorization procedure similar to that used to derive Eq. (12). In this case statistical dependencies exist between $\mathbf{a}_i$ and $\mathbf{a}_k$ for $k = i$, $k = i + 1$ and $k = i - 1$. Taking these conditions into account, the expected intensity patterns can be arranged as

$$\begin{aligned}&\langle|\mathbf{A}(f_x)|^2\rangle \\ &= |\mathbf{B}(f_x)|^2 + \sum_{i=1}^{N}(1 - |\mathbf{b}_i|^2) \\ &\quad + 2\sum_{i=1}^{N-1}\mathrm{Re}(\{\langle\exp[j(1-2\alpha)\phi_i]\rangle\langle\exp[-j(1-\alpha)\phi_{i+1}]\rangle \\ &\quad \times \langle\exp(j\alpha\phi_{i-1})\rangle - \mathbf{b}_i\mathbf{b}_{i+1}^*\}\exp(j2\pi\Delta f_x)). \qquad (14)\end{aligned}$$

The second summation in Eq. (14) would be identically zero except for the term

$$\langle\exp[j(1-2\alpha)\phi_i]\rangle \neq \langle\exp[j(1-\alpha)\phi_i]\rangle\langle\exp[-j\alpha\phi_i]\rangle$$

that arises when terms $\langle\mathbf{a}_i\mathbf{a}_{i+1}^*\rangle$ are considered. Evaluation of the expectations in Eq. (14) using the PDF of Eq. (7) in Eq. (5), together with additional trigonometric identities, leads to

$$\begin{aligned}&\langle|\mathbf{A}(f_x)|^2\rangle \\ &= |\mathbf{B}(f_x)|^2 + \sum_{i=1}^{N}(1 - |\mathbf{b}_i|^2) + 2\,\mathrm{Re}\left\{\sum_{i=1}^{N}\tan[(1-\alpha)\nu_i/2]\right. \\ &\quad \left.\times \tan(\alpha\nu_i/2)\mathbf{b}_i\mathbf{b}_{i+1}^*\exp(j2\pi\Delta f_x)\right\}. \qquad (15)\end{aligned}$$

Note that Eq. (15) is of the form

$$\langle|\mathbf{A}(f_x)|^2\rangle = |\mathbf{B}(f_x)|^2 + C_1 + C_2\cos(2\pi\Delta f_x + \Phi), \quad (16)$$

where C_1 represents the first summation in Eq. (15) and C_2 represents the magnitude of the second summation in Eq. (15). The second summation reduces to a single cosine component of phase shift Φ. Comparing Eq. (16) with Eq. (12) shows that the desired diffraction pattern $|\mathbf{A}_c(f_x)|^2$ is distorted into $|\mathbf{B}(f_x)|^2$, and the noise background changes from white to colored. It is interesting to note that, similarly to the summation term in Eq. (12), the white noise term C_1 indicates the amount of energy scattered into the noise background.[8] This result follows from the fact that the colored-noise term $C_2\cos(2\pi\Delta f_x + \Phi)$ has a period of $1/\Delta$, which is the nonredundant bandwidth of the diffraction pattern. Thus this term integrates to zero energy over the bandwidth of any given diffraction order.

5. EVALUATION OF PERFORMANCE CHANGES DUE TO PHASE BLURRING OF PSEUDORANDOM ENCODING

The results and relationships derived in Section 4 are used to quantify performance changes as a function of the blurring parameter α. These computer-simulated results are compared with experimental results found with use of an electrically addressed SLM. The possibility of using predistortion to compensate for phase distortion is also considered in the simulations.

A. Simulation and Measurement of Nearest-Neighbor Blurring

The modulation pattern is designed to produce an off-axis array of 7 × 7 spots. Without blurring the modulation is identical to the one used with the LCLV in Sections 2 and 3. The modulation consists of 128 × 128 pixels complex modulation encoded to phase by the PRE method of Subsection 4.A. The encoded phase is then blurred along the diagonal direction so that the blurred phase is $(1 - \alpha)\phi_{i,i} + \alpha\phi_{i-1,i-1}$. The resulting complex modulation pattern is then zero padded to produce a 512 × 512 array of samples. The sample array is then Fourier transformed by the FFT. A resulting intensity image is shown in Figs. 3(a) and 3(b) without and with blurring, respectively. The location of the various spots makes it qualitatively similar to the simulated image in Fig. 1(c). Metrics for various degrees of blurring (thick curves) are presented in Fig. 4. For $\alpha = 0.2$ there is substantial loss of performance in NU, and for the other two metrics the performance is substantially reduced for α as small as 0.1.

The simulated far-field intensity pattern resulting from this analysis is compared with measured intensity patterns from an electrically addressed SLM that is programmed to produce the identical blurred phase. The SLM used is a 128 × 128 pixel nematic liquid-crystal SLM from Boulder Nonlinear Systems (BNS). Additional technical specifications and experimental measurements of the device are presented in Ref. 10. One key property of this SLM is that intensity and interferometric images of the SLM indicate no noticeable coupling between nearest-neighbor pixels. Apparently the grounding electrodes between the pixels significantly reduce fringing fields compared with those found in the LCLV. As with the LCLV, the on-axis light is usually much brighter than the spot array. This is due not to blurring but rather to the relatively large percentage of light that is reflected from the cover glass of the SLM.

The experimental diffraction patterns produced for no blurring and for blurring ($\alpha = 0.5$) are presented in Figs. 3(d) and 3(e), respectively. Except for the on-axis spot that is due to reflections from the cover glass, these diffraction patterns are qualitatively similar to the simu-

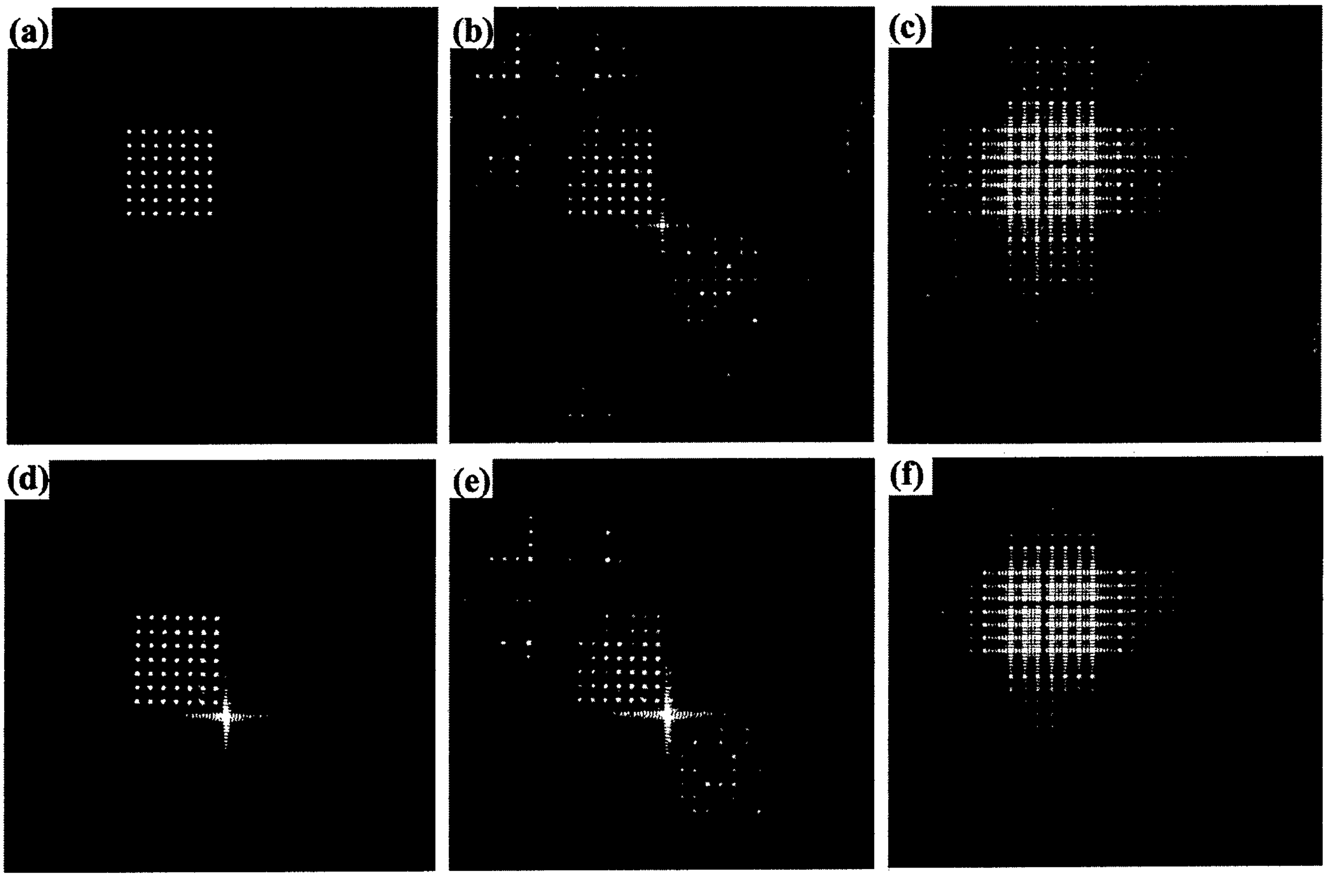


Fig. 3. Far-field diffraction patterns resulting from the identical PRE design of a spot array generator: (a) and (d) without phase blurring, (b) and (e) with phase blurring $\alpha = 0.5$, and (c) and (f) with phase blurring of a predistorted phase for $\alpha = 0.5$. (a) and (b) are simulated, (c) is the average of ten simulations each using a different random sequence for PRE, (d) and (e) are as measured for the BNS SLM, and (f) is the expected far-field intensity pattern as calculated with Eq. (15). Maximum white in the gray-scale images corresponds in (a) and (d) to 30%, in (b) and (e) to 50%, and in (c) and (f) to 1% of the average peak intensity of the desired spot array.

lated patterns in Figs. 3(a) and 3(b). The metrics calculated from the SLM diffraction patterns are plotted as dots in Fig. 4. Considering the inaccuracies in programming coherent SLM's and in measuring coherent optical patterns, the experimental diffraction patterns produce performance metrics that are reasonably similar to the simulated performance metrics. For SPR in Fig. 4 the discrepancy is quite noticeable for $\alpha < 0.06$. In this range the brightest noise spot is due to speckle noise. However, for larger values of α the peak noise is from peaks in the first harmonic of the spot array. (The on-axis spot is omitted from consideration in both experimental and simulated SPR, as discussed further in Appendix A.) The experimental measurements suggest that speckle noise is higher in practice than for the simulations. The higher speckle noise accounts in part for the measured values of NU being higher than the simulated values. Another influence on NU could be multiple coherent reflections between the SLM and its cover glass.[17] The magnitude of the discrepancy between NU as measured and as simulated is in line with previous measurements of NU.[10,17] Also, for small values of α the denominator term E_{-1} of E_1/E_{-1} is dominated by speckle noise. Thus in this range E_1/E_{-1} measures the SNR in the vicinity of the mirror (i.e., the -1) order.

Figure 4 also shows that a slight degree of blurring can improve individual metrics. This is not unreasonable. For example, introducing (even a small degree of) phase blurring into a diffuser is known to convert its far-field intensity statistics from exponentially distributed into a modified Rician, which is much less likely to produce as large a maximum-intensity peak as does the exponential.[5]

B. Simulated Correction of Blurring by Predistortion of Phase and Limitations

The effect of phase blurring can be compensated by convolving the encoded phase modulation ϕ_i with the inverse filter h_i^{-1}. The inverse filter is defined such that convolving h_i with its inverse produces the delta function $(h_i * h_i^{-1} = \delta_i)$. Therefore it is possible to numerically compensate the phase by first predistorting phase $(h_i^{-1} * \phi_i)$ and then blurring it $(h_i * h_i^{-1} * \phi_i = \phi_i)$. For the discretely sampled SLM with a phase PSF that is a discrete version of Eq. (1), $h_1 = (1 - \alpha)\delta_i + \alpha\delta_{i-1}$, the inverse filter is known to be

$$h_i^{-1} = \frac{1}{1-\alpha}\left(\frac{-\alpha}{1-\alpha}\right)^i; \quad i \geqslant 0. \tag{17}$$

This filter is stable for $\alpha < 0.5$ and marginally stable for $\alpha = 0.5$. For $1 > \alpha > 0.5$ the data can be filtered in the reverse (anticausal) direction to ensure a numerically stable solution.[18,19] It is clear from this discussion that the predistorted phase can be exactly compensated by the inverse filter in Eq. (17) *in a numerical simulation.*

However, there is a practical limitation to using predistortion that can be seen by examining Eq. (17). The problem is that the predistorted phase range can be much greater than 2π. This can be appreciated by convolving the inverse filter with a step function of height 2π. For $\alpha = 0.5$ the sequence of predistorted phases is $[4\pi, 0, 4\pi, 0, \ldots]$, and for $\alpha = 0.3$ the predistorted phase sequence is $[2.8\pi, 1.64\pi, 2.16\pi, 1.94\pi, \ldots]$. The predistorted phase range is even greater for the PRE-designed 7×7 spot array generator. We find that the total phase range is 96π for $\alpha = 0.5$, 7.6π for $\alpha = 0.45$, and 5.8π for $\alpha = 0.3$. Since LCLV's are limited in phase range to near 2π (for practical reasons) it is likely that they will not be able to respond to predistorted addressing signals corresponding to phase shifts that are greatly in excess of 2π. Also, modding of the predistorted phase into the 2π range of the LCLV is not an acceptable option, and it will not correctly compensate for phase blurring.

There is also a limitation to using a blur-free, electrically addressed SLM to demonstrate phase compensation experimentally. Since the phase is both blurred and predistorted in the attached computer, the signal applied to the SLM is exactly the same signal as would have been applied to the SLM if there were no phase blurring.

Examination of Eq. (13) suggests that one could apply the inverse filter to the desired phase ψ_i, which also is blurred by the single-lag phase PSF. Various simulations indicate that this phase distortion contributes to the appearance of undesired diffraction orders to a much greater degree than do the amplitude distortions of Eq. (13). While both amplitude and phase distortions can be exactly compensated for by predistorting ϕ_i, it is interesting to consider to what degree that predistortion of the desired phase ψ_i reduces the effects of blurring. The resulting far-field patterns are shown in Figs. 3(c) and 3(f). The performance in Fig. 4 for predistorted phase (thin curves) is generally better than for phase that is not predistorted (thick curves). For SPR and NU the performance is close to that for the original design (i.e., for $\alpha = 0$). The metric E_1/E_{-1} (inset in Fig. 4) is much larger after predistortion of the phase. This change is due to the following: (1) The contribution to E_{-1} from harmonics is negligible, as can be evaluated from the first term of Eq. (16). (2) Essentially only the speckle pattern contributes energy to E_{-1}. (3) The speckle pattern intensity is weakest in the region of the -1 order, and it is much weaker than for the case of no predistortion. For example, for $\alpha = 0.5$, C_1, and C_2 in Eq. (16) are of similar magnitude for the predistorted design, and the -1 order is located in a region where the cosine term subtracts from and nearly cancels C_1. The cosine variation of the speckle noise can be seen both for the simulated [Fig. 3(c)] and for the theoretical [Fig. 3(f)] expected intensity pattern. Specifically, the simulated expectation is the average of ten realizations of the identical PRE design, each using a different uncorrelated realization of the random sequence s_i, and the theory is the expectation from Eq. (15). The average of several realizations makes it easier to see and compare low-level features with the theory than does a single realization.

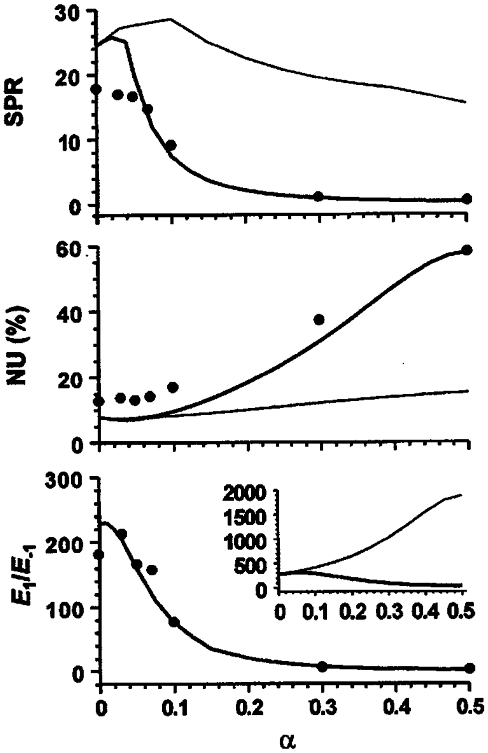


Fig. 4. Performance of PRE as a function of α, the degree of phase blurring. Solid curves report the computer-simulated results for the effect of phase blurring. Dots show the experimentally measured values obtained with the BNS SLM. Thin curves report the simulated performance if the phase is first predistorted through the inverse filter of Eq. (17). The inset plots E_1/E_{-1} over an extended range that shows the simulated performance for phase blurring and for phase blurring of the predistorted phase.

Two interesting results come out of this simulation. First, the compensation of the desired phase ψ_i in Eq. (13) results in values of NU and SPR that are close to the values for no phase blurring. Second, the compensated diffraction patterns appear very similar to the diffraction pattern without blurring. The main difference is that the speckle pattern in the compensated diffraction pattern [Figs. 3(c) and 3(f)] has a noticeable sinusoidal component ($C_2/C_1 \approx 0.6$, not just for $\alpha = 0.5$ but for $0.3 \leqslant \alpha \leqslant 0.5$), while the speckle pattern in diffraction patterns without burring have no sinusoidal component, as indicated in Eq. (12). This is especially interesting because the sinusoidal component of speckle for the phase-blurred modulation in Figs. 3(b) and 3(e) is not apparent (since $C_2/C_1 < 0.01$).

Finding an exact inverse filter for a two-dimensional filter is nontrivial, and some approximation is usually required.[18,19] For the 7×7 LCLV phase PSF measured in Section 3 we computed an approximate inverse by using the discrete Fourier transform to calculate the two-dimensional phase spectrum $H(f_x, f_y)$. The exact inverse filter is $H^{-1}(f_x, f_y)$. However, the function contains singularities. For this reason, amplitudes in excess of a threshold value γ are set to γ so that for $|H^{-1}| > \gamma$, the modified filter is $\gamma \exp[\arg(H^{-1})]$. We used the modified inverse filter to predistort the phase ϕ_i of the spot array generator design. We found through repeated experiments that a value of γ that is 25× larger than the minimum value of $|H^{-1}|$ reduces the effects of blurring on the performance metrics the most (as reported in the last column of Table 2). While predistortion of the blurred phase significantly corrects for blurring, approximations inherent in the inverse filter method do not completely restore the performance to the levels achieved if blurring were not present. As with the single-lag blurring function, the predistorted phase range (which is 40π) is too large to experimentally implement. Even for an isolated 2π step the predistorted phase range is 6.9π. Furthermore, owing to incomplete compensation, the phase range of the corrected step increases from 2π to 2.3π.

These results show that predistortion is numerically possible but physically quite difficult owing to the limited phase range of most LCLV's. These simulations do provide insight into the effect of blurring. In particular, the compensation of the desired phase ψ_i provides an example of a significant spatial variation of the speckle noise background. Also, these analyses show that modding of the phase into a 2π range, which often is taken for granted in CGH design, is not necessarily possible because of the nonlinearity inherent in phase blurring. This limitation is considered further in Subsections 5.C and 5.D.

C. Evaluation of Blurring on a Linear Phase Ramp

A standard method of evaluating phase distortion is to consider the effect on a single frequency f_0.[7] This corresponds to a desired phase that is a linear phase ramp or $\psi(x) = 2\pi f_0 x$. On the basis of current SLM's we also will assume that the phase is modded into a 2π range. The modded phase is a periodic function of period $1/f_0$. For the phase PSF of Eq. (1) the blurred phase over one period can be written as

$$\psi_b(x) = \begin{cases} 2\pi[(x - \alpha\Delta)f_0 + \alpha] & \text{if } 0 \leqslant x < \Delta \\ 2\pi(x - \alpha\Delta)f_0 & \text{if } \Delta \leqslant x < 1/f_0 \end{cases}. \tag{18}$$

The blurred phase for $\Delta f_0 = 1/2$ and $\alpha = 1/3$ is plotted in Fig. 5. It can be seen that blurring reduces the phase range (thus introducing a dc component) and causes a phase discontinuity (thus producing higher-frequency components). The strengths of the desired and undesired diffraction orders are found by evaluating the integral

$$d_k = f_0 \int_0^{1/f_0} \exp[j\psi_b(x)]\exp(-j2\pi k f_0 x)\mathrm{d}x, \tag{19}$$

which are the Fourier series coefficients of the complex modulation. The intensity of the diffraction orders evaluates to

$$|d_k|^2 = \begin{cases} (1 - \Delta f_0)^2 + (\Delta f_0)^2 + 2(\Delta f_0)(1 - \Delta f_0)\cos(2\pi\alpha) \\ \quad \text{if} \quad k = 1 \\ |[2/\pi(1 - k)]\sin[\pi\Delta f_0(1 - k)]\sin(\pi\alpha)|^2 \\ \quad \text{if} \quad k \neq 1 \end{cases}. \tag{20}$$

Thus blurring applied to an ideal blazed grating introduces undesired diffraction orders at frequencies kf_0, $k \neq 1$. However, if the linear phase were a continuous unmodded ramp, then the desired phase and blurred-phase patterns would both produce a single diffraction order at f_0. The spots would be identical in intensity and differ only by a constant phase shift $2\pi\alpha\Delta f_0$.

The validity of Eq. (20) was compared with a computer simulation and experimental implementation with use of the BNS SLM. A comparison of the results for two values of α are summarized in Table 3. For the computer simulation the Fourier series coefficients are calculated by taking the FFT of one period of the blurred modulation.[20] One period in the simulation consists of 4.75Δ or, equivalently, $\Delta f_0 = 4/19$. Equation (20) and the simulation give nearly identical results, indicating that Eq. (20) is correct. This phase modulation also is programmed on the 128 × 128 pixel SLM. The simulated and experimental diffraction patterns are compared in Table 3. The results are reasonably similar to the theory in relative strength. This suggests that the SLM produces a phase modulation that is similar to the modeled phase.

D. Comparison of the Phase-Ramp Model with the Pseudorandom-Encoding Model of Blurring

In this section we compare the effect of phase blurring on PRE as modeled in Eq. (15) with a traditional diffraction-efficiency model based on Eq. (20).

Models of periodic modulation similar to the one in Subsection 5.C have been widely used to predict the diffraction efficiency of DOE's. These models are based on the assumption that the desired diffraction pattern recon-

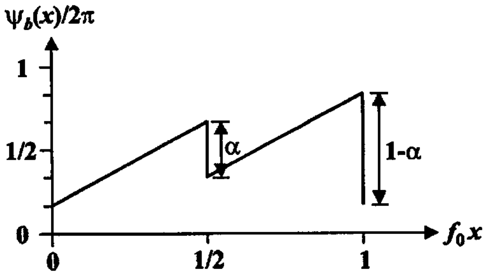


Fig. 5. Illustration of the distortion of a modded phase ramp of 2π range due to the phase PSF of Eq. (1).

Table 3. Diffraction-Order Intensities for a Modded Phase Ramp That Is Phase Blurred[a]

Method of Analysis	I_{-1}	I_0	I_1	I_2	I_3
$\propto = 0.3$					
Theory, Eq. (20)	0.106	0.165	1	0.165	0.106
Simulation	0.102	0.162	1	0.148	0.102
Experiment	0.150	—	1	0.136	0.212
$\alpha = 0.5$					
Theory, Eq. (20)	0.266	0.414	1	0.414	0.266
Simulation	0.256	0.405	1	0.368	0.253
Experiment	0.134	—	1	0.566	0.380

[a] For $\Delta f_0 = 4/19$. Data are normalized to I_1. $\propto$ is the degree of phase blurring.

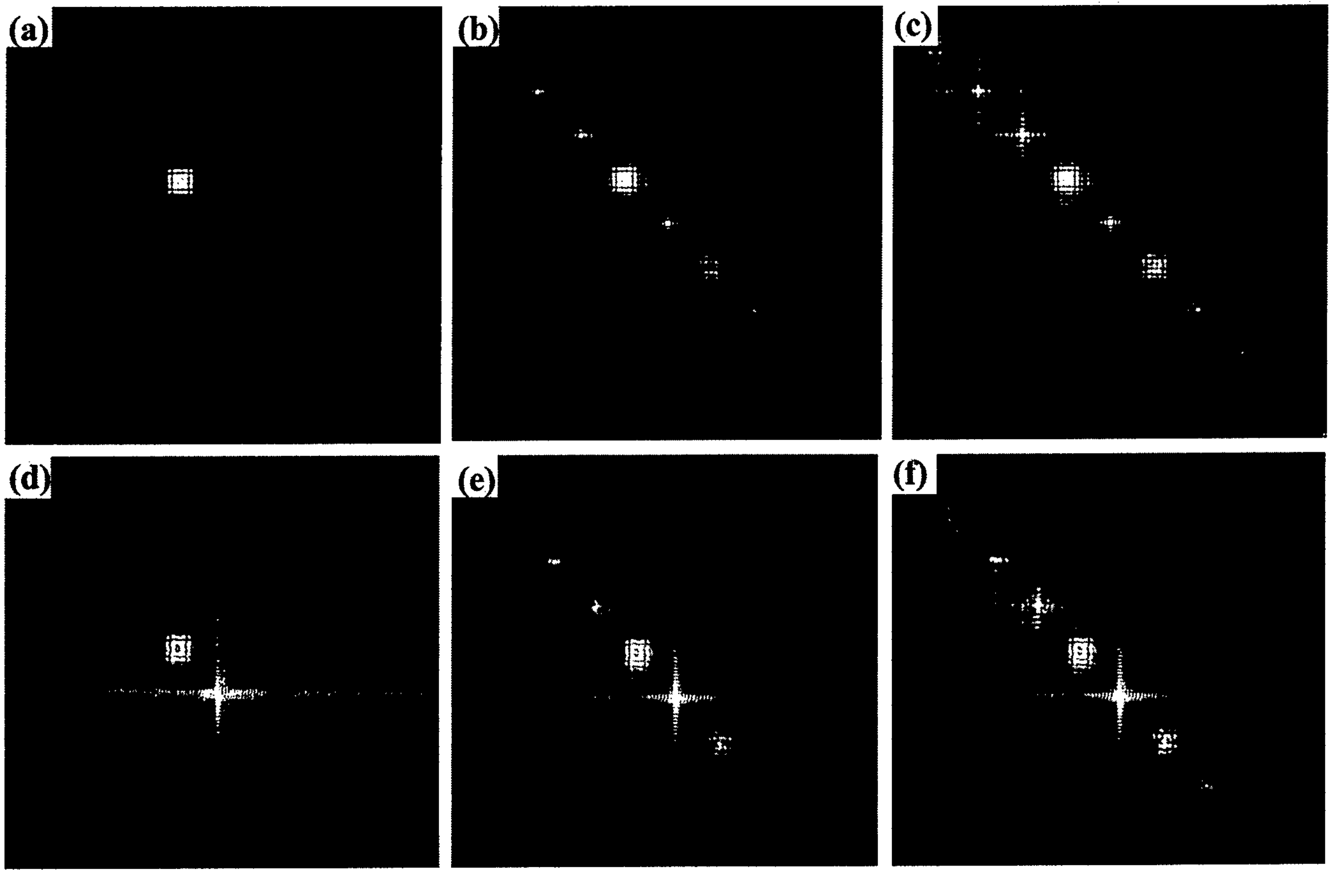


Fig. 6. Far-field diffraction patterns resulting from the identical PRE design of a spot array generator: (a) and (d) without phase blurring, (b) and (e) with phase blurring $\alpha = 0.3$, and (c) and (f) with phase blurring $\alpha = 0.5$. (a)–(c) are simulated, and (d)–(f) are as measured for the BNS SLM. Maximum white in the gray-scale images corresponds to 15% of the peak intensity level of the desired spot array.

Table 4. Diffraction-Order Intensities That Result from Phase Blurring of PRE[a]

Method of Analysis	E_{-1}	E_0	E_1	E_2	E_3
$\propto = 0.3$					
Eq. (20)	0.106	0.165	1	0.165	0.106
Theory, Eq. (15)	0.122	0.200	1	0.200	0.144
Simulation	0.127	0.211	1	0.202	0.146
Experiment	0.131	—	1	0.228	0.148
$\alpha = 0.5$					
Eq. (20)	0.266	0.414	1	0.414	0.266
Theory, Eq. (15)	0.278	0.455	1	0.416	0.305
Simulation	0.285	0.463	1	0.406	0.305
Experiment	0.258	—	1	0.555	0.284

[a] For $\Delta f_0 = 4/19$. Data are normalized to E_1. $\propto$ is the degree of phase blurring.

structs around a specific diffraction order. In this paper we have been considering the desired order to be the $k = 1$ order at spatial frequency f_0. This center frequency can be viewed as the carrier frequency for the desired signal. In some cases (particularly if the modulation on the carrier has a small bandwidth and the modulation depth of the carrier is small) the diffraction efficiency for the modded linear phase ramp [found from Eq. (20)] should correspond closely to the efficiency for the signal-modulated carrier. However, for the designs based on PRE (which ideally do not produce additional harmonics at kf_0, $k \neq 1$) the bandwidth is greater than f_0 and the deviation between the phase ramp and the PRE phase modulation is frequently as large as $\pm \pi$.

To illustrate the differences between the two models we define a desired function that has a much smaller bandwidth. This is achieved by sampling the desired phase function for the 7 × 7 spot array generator more finely. The resulting function consists of 2 × 2 rather than 4 × 4 unit cells, and the spacing between the resulting spots is reduced by a factor of 1/2. A phase ramp is added to the desired modulation so that the spot array is centered $\Delta f_0 = 4/19$, the same frequency used in Subsection 5.C. The simulated diffraction pattern calculated from the 512 × 512 sample FFT of the blurred PRE design is shown in Figs. 6(a)–6(c). Compared with the previous design in Figs. 1 and 3, the desired diffraction pattern is much more separated from the harmonic patterns. However, there is still some overlap between the orders. Based on the shape of each order a unique and nonoverlapping window is chosen for performing the integration of intensity. The window for the 1 and -1 orders is chosen to be a cross-shaped area composed of two rectangles oriented at right angles to each other in x and y. Each rectangle is 33 × 65 samples. The window for the 0 and

3 orders is also a symmetric cross with arms in x and y. Each rectangle is 17 × 201 samples. The window for the 2 order is a rectangle that is 83 × 77 samples. Identical windows are applied to the simulation of the blurred PRE (i.e., the FFT of encoded modulation) and to the model [i.e., the expected intensity calculated from Eq. 15)]. The results are compared in Table 4 and are found to be nearly identical. These results do not compare quite as closely with the diffraction-order strengths calculated from Eq. (20). In particular, the intensities for Eq.(20) are symmetric around $k = 1$, whereas the energies calculated from Eq. (15) are not symmetric.

Perhaps even closer correspondence would be obtained if it were possible to further increase the carrier frequency and the separation between the orders. However, Eq. (15) accurately predicts the energy in a given window despite the overlap. Additionally, unlike Eq. (20), the simulations [Figs. 6(a)–6(c)] or Eq. (15) provide detailed intensity patterns around each harmonic frequency. Most of the same features seen in the simulations are also seen in Figs. 6(d)–6(f), which are the corresponding experimental diffraction patterns from the SLM. The measured values of energy E_k in Table 4 are somewhat closer to the theory with use of Eq. (15), with the greatest discrepancy (~25%) being for E_2 for $\alpha = 0.5$.

The model of blurring of PRE [Eq. (15)] appears to provide additional accuracy compared with the basic linear model of Eq. (20). The ability of the model to predict the detailed effects of blurring when the unwanted harmonic patterns overlap with the desired designs (such as in Fig. 3) is the key advantage of the blurred PRE model.

6. SUMMARY AND CONCLUSIONS

The historical approach of describing the resolution of LCLV's as a function of input level may not be the most appropriate for Fourier transform applications such as beam steering, pattern generation and matched filtering. The phase range desirable for these applications is near 2π and over this range the resolution of most LCLV's used as display devices changes dramatically. However, as measured for a specific LCLV, the phase resolution does not change noticeably, which suggests the suitability of the space-invariant phase-blurring model. Using the experimentally measured phase PSF in simulations produces far-field patterns with distortion products that are quite similar to those actually measured from the LCLV. Simulations in which the phase PSF is varied in diameter show that the blur diameter needs to be a very small fraction of the pixel spacing for the effects of blurring to be negligible for the 7 × 7 spot array generator design.

Additional experiments are performed with an electrically addressed SLM that has no appreciable coupling between nearest-neighbor pixels. Filtering of the electrical address signal is used to experimentally introduce various degrees of phase blurring. Performance is dramatically reduced for $\alpha = 0.1$, which, as with the LCLV experiments, indicates that a small degree of phase blurring can significantly alter performance. The effects of phase blurring on PRE designs can be compensated to a large extent by applying an inverse filter to the phase. However, this approach is not practical since it requires that the SLM have a phase range that greatly exceeds 2π.

A Fourier series analysis of a modded phase ramp subjected to phase blurring is performed to provide a clear example of the nonlinear generation of harmonics that is due to blurring. However, this result does not accurately predict the energy found in the diffraction orders of the PRE-designed spot array. Much more detailed information can be found from the closed-form expression Eq. (15) for the expected far-field intensity pattern. This expression predicts the distortion of each effective pixel value, which leads to the generation of undesired harmonics. It also predicts the expected noise power spectrum, which, owing to phase blurring, has a nonwhite distribution. Equation (15) is no more numerically efficient than numerical simulations involving the application of the PRE algorithm, blurring of the phase encoding, and calculation of the Fourier transform. However, the recognition that Eq. (15) is of the form of Eq. (16) provides insight into the nonlinear effects of phase blurring and a base from which to develop models of various performance metrics.

APPENDIX A

The definition and method of calculating the performance metrics used in this paper are collected here for easy reference. The metrics of signal-to-noise ratio (SNR), signal-to-peak-noise ratio (SPR), nonuniformity (NU), and diffraction efficiency (η) are reviewed from earlier work.[5,8–10] Additional metrics that are useful for describing the nonlinear effects of blurring are the intensities I_k and the energies E_k of the kth diffraction order that result from blurring of the desired modulation.

The orders are defined so that $k = 0$ corresponds to light on the optical axis and $k = 1$ corresponds to the desired reconstruction centered at frequency f_0. The energy in a spot array and its harmonic orders is compared as a ratio. A particular window of integration around each order is chosen to make a fair comparison between the values of E_k. The same window is used for comparing experiment, simulation, and theory. We present this data as a ratio of energies such as E_k/E_1. For most measurements reported in all sections except Subsections 5.B and 5.D, we calculate E_1 so as to minimize energy contributions from speckle. This is accomplished by summing the intensities only at the centers of the 49 desired spots. The ratio E_{-1}/E_1 is calculated by summing the intensities at the same 49 frequencies, appropriately centered around the frequency $-f_0$. Modified windows are used to minimize the variability of speckle from the -1 order and the overlap between harmonics. These windows are specified in Subsections 5.B and 5.D, respectively.

In this paper the ratio of E_1 to either E_0 or E_{-1} is reported to show the production of unwanted orders by blurring. We do not report experimental measurements of E_1/E_0, because reflections from the cover glass of both SLM's produce bright spots on the optical axis that are not accounted for in the theory and simulations. Rather than adjusting the theory to the specific modulators, we chose to experimentally measure E_1/E_{-1}. For this same reason we do not report experimentally measured values of I_0 and η.

The SNR is the average spot intensity (specifically, $E_1/49$ for the 7×7 spot array) divided by the average noise intensity. The noise intensity is calculated from the two quadrants in the diffraction other than where the desired diffraction signal and its harmonics appear. The identical region is used for calculations from the simulations and the experiments. The SPR is the average spot intensity divided by the most intense noise peak found in the diffraction pattern. The on-axis spot and the sidelobes of the spots in the vicinity of the desired spot array are omitted from this calculation. However, spots from the harmonic orders are included in the calculation of SPR. The NU is the standard deviation of the desired spot intensities relative to the average spot intensity. The standard deviation and average are calculated from the peak intensities of the 49 spots located around f_0. The diffraction efficiency values are reported here for purposes of making relative comparisons of energy distribution between desired and undesired portions of the light distribution. The value of η is calculated in the simulations by summing the energy in a window around the 7×7 spot array and dividing by the total energy of the FFT, which cover only the frequency range of the SLM grating order $1/\Delta$. Energy that appears in adjacent grating orders, which depends on the pixel aperture, is not considered in the theory or simulations of the diffraction patterns. As mentioned above, experimental measurements of diffraction efficiency are not reported because the presence of the reflection from the cover glass obscures the measurement of I_0.

ACKNOWLEDGMENTS

This study was supported by U.S. Office of Naval Research grant N00014-96-1-1296 and NASA cooperative agreement NCC5-222. The diffractive optical elements were fabricated by Honeywell as part of the Defense Advance Research Projects Agency DOE foundry run.

*When the work was performed, the authors were with University of Louisville. M. Duelli's current address is Optical Coating Laboratory, Inc., 2789 Northpoint Parkway, MS 125-3, Santa Rosa, California 95407-7397.

Address correspondence to R. W. Cohn at the location on the title page or by e-mail: rwcohn@louisville.edu.

REFERENCES

1. D. Casasent, "Performance evaluation of spatial light modulators," Appl. Opt. **18**, 2445–2453 (1979).
2. A. D. Fisher and J. N. Lee, "The current status of two-dimensional spatial light modulator technology," in *Optical and Hybrid Computing*, H. H. Szu, ed., Proc. SPIE **634**, 352–371 (1986).
3. T. D. Hudson and D. A. Gregory, "Optically addressed spatial light modulators," Opt. Laser Technol. **23**, 297–302 (1991).
4. D. V. Wick, T. Martinez, M. V. Wood, J. M. Wilkes, M. T. Gruneisen, V. L. Berenberg, M. V. Vasil'ev, A. P. Onokhov, and L. A. Beresnev, "Deformed-helix ferroelectric liquid-crystal spatial light modulator that demonstrates high diffraction efficiency and 370-line pairs/mm resolution," Appl. Opt. **38**, 3798–3803 (1999).
5. R. W. Cohn, A. A. Vasiliev, W. Liu, and D. L. Hill, "Fully complex diffractive optics by means of patterned diffuser arrays: encoding concept and implications for fabrication," J. Opt. Soc. Am. A **14**, 1110–1123 (1997).
6. R. W. Cohn and L. G. Hassebrook, "Representations of fully complex functions on real-time spatial light modulators," in *Optical Information Processing*, F. T. S. Yu and S. Jutamulia, eds. (Cambridge U. Press, Cambridge, UK, 1998), Chap. 15, pp. 396–432.
7. W. J. Dallas, "Computer-generated holograms," in *The Computer in Optical Research*, B. R. Frieden, ed. (Springer-Verlag, Berlin, 1980), Chap. 6, pp. 291–366.
8. R. W. Cohn and M. Liang, "Pseudorandom phase-only encoding of real-time spatial light modulators," Appl. Opt. **35**, 2488–2497 (1996).
9. R. W. Cohn and M. Duelli, "Ternary pseudorandom encoding of Fourier transform holograms," J. Opt. Soc. Am. A **16**, 71–84; "errata," 1089–1090 (1999).
10. M. Duelli, M. Reece, and R. W. Cohn, "A modified minimum distance criterion for blended random and nonrandom encoding," J. Opt. Soc. Am. A **16**, 2425–2438 (1999).
11. R. W. Cohn and M. Liang, "Approximating fully complex spatial modulation with pseudorandom phase-only modulation," Appl. Opt. **33**, 4406–4415 (1994).
12. G. Paul-Hus and Y. Sheng, "Optical on-axis real-time phase-dominant correlator using liquid crystal television," Opt. Eng. **32**, 2165–2172 (1993).
13. R. W. Cohn and J. L. Horner, "Effects of systematic phase errors on phase-only correlation," Appl. Opt. **33**, 5432–5439 (1994).
14. E. Shafir, H. Bernstein, A. A. Friesem, and H. Grubel, "Method for measuring the spatial frequency response of phase-modulating spatial light modulators," Opt. Eng. **27**, 71–74 (1988).
15. L. G. Hassebrook, M. E. Lhamon, R. C. Daley, R. W. Cohn, and M. Liang, "Random phase encoding of composite fully-complex filters," Opt. Lett. **21**, 272–274 (1996).
16. R. W. Cohn, "Analyzing the encoding range of amplitude-phase coupled spatial light modulators," Opt. Eng. **38**, 361–367 (1999).
17. M. Duelli, D. L. Hill, and R. W. Cohn, "Frequency swept measurements of coherent diffraction patterns," Appl. Opt. **37**, 8131–8133 (1998).
18. R. R. Read, J. L. Shanks, and S. Treitel, "Two dimensional recursive filtering," in *Picture Processing and Digital Filtering*, Vol. 6 of Topics in Applied Physics, T. S. Huang, ed. (Springer-Verlag, Berlin, 1979), pp. 131–176.
19. J. S. Lim, "Image restoration," in *Two-Dimensional Signal and Image Processing* (Prentice-Hall, Englewood Cliffs, N.J., 1990), pp. 524–588.
20. A. V. Oppenheim and R. W. Shafer, *Discrete-Time Signal Processing* (Prentice-Hall, Englewood Cliffs, N.J., 1989), p. 530.

High-diffraction-efficiency pseudorandom encoding

Yongyi Yang, Henry Stark, and Damla Gurkan

Department of Electrical and Computer Engineering, Illinois Institute of Technology, 3301 South Dearborn Street, Chicago, Illinois 60616

Christy L. Lawson and Robert W. Cohn

The ElectroOptics Research Institute, University of Louisville, Louisville, Kentucky 40292

Received February 9, 1999; revised manuscript received June 21, 1999; accepted July 15, 1999

Pseudorandom encoding (PRE) is a statistics-based procedure in which a pure-phase spatial light modulator (SLM) can yield, on the average, the prescribed diffraction pattern specified by the user. We seek to combine PRE with the optimization of an aperture-based target function. The target function is a fully complex input transmittance, unrealizable by a phase-only SLM, that generates a prescribed light intensity. The optimization is done to increase the diffraction efficiency of the overall process. We compare three optimization methods—Monte Carlo simulation, a genetic algorithm, and a gradient search—for maximizing the diffraction efficiency of a spot-array generator. Calculated solutions are then encoded by PRE, and the resulting diffraction patterns are computer simulated. Details on the complexity of each procedure are furnished, as well as comparisons on the quality, such as uniformity of the output spot array. © 2000 Optical Society of America [S0740-3232(00)01002-4]

OCIS codes: 050.1970, 090.1970, 220.4830, 230.6120.

1. INTRODUCTION

Pseudorandom encoding (PRE) is a procedure that enables a pure-phase spatial light modulator (SLM) to approximately produce the same Fraunhofer diffraction intensity that would result from a desired, but unrealizable, fully complex filter. By a fully complex filter, we mean a device in which *both* amplitude and phase can be varied to produce the desired diffraction pattern.

There is a sizable literature on the design of input generating functions that achieve prescribed far-field intensities subject to a phase-only (PO) constraint.[1–25] Indeed, PRE is only one of many design methods that exist for this purpose. The iterative Fourier transform algorithm is known to give excellent results[4–7,9,10] but requires at least two Fourier transforms per iteration in a procedure that, depending on the complexity of the prescribed diffraction intensity, may extend to hundreds of iterations. Simulated annealing, another powerful method, is likely to yield a global optimum in a phase optimization problem but is reported to be slow[26,27] and thus may be unsuitable for real-time systems that require adaptive redesign of the SLM's modulation pattern.[25] For such systems a noniterative encoding algorithm that requires only a few numerical operations per SLM pixel would seem to be the preferred choice. PRE is such an encoding procedure. However, the overall diffraction efficiency of the process must be reasonably high. If some sort of optimization is required to increase the diffraction efficiency, the additive time component associated with the optimization algorithm becomes a major factor in evaluating its efficiency.

For readers not familiar with PRE, we furnish a brief review in Appendix A. More extensive discussions of PRE appear in Refs. 23–25. Some examples of design by the iterative Fourier transform algorithm are given in Refs. 4–6. Iterative Fourier transform algorithm design from a vector-space point of view is discussed in Refs. 9 and 10. A tutorial discussion of vector-space methods is found in Ref. 28. General approaches to diffractive-optics design are found in Refs. 1 and 2. A discussion of virtual source arrays is furnished in Refs. 20 and 21. Various optimation methods useful in diffractive-optics design can be found in Refs. 20, 22, 26, and 27.

2. DIFFRACTION EFFICIENCY AND SIGNAL-TO-NOISE RATIO

The diffraction efficiency of interest here is the so-called input diffraction efficiency η_{in}, defined by (for simplicity, we use only one-dimensional notation)

$$\eta_{\text{in}} = \begin{cases} \dfrac{1}{L}\displaystyle\int_{-L/2}^{L/2} |g(x)|^2\,\mathrm{d}x & \text{(continuous)} \quad (1) \\ \dfrac{1}{N}\displaystyle\sum_{i=1}^{N} |g(i\Delta)|^2 & \text{(discrete)} \quad (2) \end{cases},$$

where L is the aperture dimension, $g(x)$ is the aperture function that generates the desired far field $G(u)$, i.e., $g(x) \leftrightarrow G(u)$, N is the number of discrete cells, and Δ is the cell size, i.e., $\Delta = L/N$.

A PO array of cells can be modeled by the transmittance

$$g_1(x) = K\sum_{i=1}^{N} \exp[j\phi(i\Delta)]\mathrm{rect}\left(\frac{x - i\Delta - \Delta/2}{\Delta}\right), \qquad (3)$$

where $\phi(i\Delta)$ is the phase of the ith cell and K is a normalizing constant of no interest at present and is therefore set equal to unity. In PRE the phases $\phi(i\Delta)$ are treated as random variables whose statistics are to be adjusted to produce an approximation to the prescribed diffraction pattern. Replacing $g_1(x)$ with $g_1(n\Delta)$ for convenience, we compute the far-field amplitude, resulting from the uniform illumination of $g_1(x)$, as

$$G_1(u) = \Delta\sum_{n=1}^{N} \exp[j\phi(n\Delta)]\exp(-j2\pi un\Delta) \qquad (4)$$

and its expected value as

$$\overline{G_1(u)} = \Delta\sum_{n=1}^{N} g(n\Delta)\exp(-j2\pi un\Delta), \qquad (5)$$

where we have used the fact that $E\{\exp[j\phi(n\Delta)]\} = g(n\Delta)$ for unbiased estimation of $g(n\Delta)$.

The expected value of field fluctuations, $\sigma_1^2(u)$, is given by

$$\begin{aligned}\sigma_1^2(u) &= \overline{|G_1(u)|^2} - [\overline{G_1(u)}]^2\\ &= N - \sum_{n=1}^{N} |g(n\Delta)|^2 = N(1 - \eta_{\mathrm{in}}). \qquad (6)\end{aligned}$$

A measure of field variability at spatial frequency u is the signal-to-noise ratio (SNR), given by

$$\mathrm{SNR}_1 = \frac{[\overline{G_1(u)}]^2}{\sigma_1^2} = \frac{[\overline{G_1(u)}]^2}{N[1 - \eta_{\mathrm{in}}]}. \qquad (7)$$

Clearly, the larger the SNR is, the less is the variability that one would observe from realization to realization. Equation (7) provides a powerful incentive for raising the diffraction intensity in PRE: An increase in η_{in} from 20% to 60% doubles the SNR. We will observe this phenomenon in the numerical results.

3. LAW OF LARGE NUMBERS

Consider next an array consisting of $2N$ cells, each of width $\Delta/2$. The aperture size and the energy entering the system remain the same as those of an N-cell array with cell width Δ. In this case the far field is

$$G_2(u) = \sum_{n=1}^{2N} \exp[j\phi(n\Delta/2)]\exp(-j2\pi un\Delta/2) \qquad (8)$$

$$\begin{aligned}&= \sum_{n=1}^{N} \{\exp[j\phi((2n - 1)\Delta/2)]\\ &\quad \times \exp[-j2\pi u(2n - 1)\Delta/2]\\ &\quad + \exp[j\phi(2n\Delta/2)]\exp(-j2\pi u2n\Delta/2)\}, \qquad (9)\end{aligned}$$

and its expected value is

$$\begin{aligned}\overline{G_2(u)} = \sum_{n=1}^{N} [&g((2n - 1)\Delta/2)\exp(-j2\pi un\Delta)\\ &\times \exp(j\pi u\Delta) + g(n\Delta)\exp(-j2\pi un\Delta)]. \qquad (10)\end{aligned}$$

In regions where $\pi u\Delta \ll 1$ (the useful operating range), $\overline{G_2(u)}$ is approximately given by

$$\begin{aligned}\overline{G_2(u)} &\cong \sum_{n=1}^{N} g((n - 1/2)\Delta)\exp(-j2\pi un\Delta)\\ &\quad + \sum_{n=1}^{N} g(n\Delta)\exp(-j2\pi un\Delta)\\ &\cong 2\overline{G_1(u)}. \qquad (11)\end{aligned}$$

Likewise, the mean-square value of the noise is

$$\begin{aligned}\sigma_2^2(u) &= \overline{|G_2(u)|^2} - [\overline{G_2(u)}]^2\\ &= 2N - \sum_{n=1}^{2N} |g(n\Delta/2)|^2\\ &= 2\sigma_1^2(u). \qquad (12)\end{aligned}$$

Hence

$$\mathrm{SNR}_2 = \frac{[\overline{G_2(u)}]^2}{\sigma_2^2} \cong 2 \times \mathrm{SNR}_1. \qquad (13)$$

Thus, for the same input energy, a doubling of the SNR is achieved by doubling the number of phase cells. Thus the field variability has been reduced, and it is in this sense that the law of large numbers works for PRE.

4. OPTIMIZATION OF SPOT ARRAYS

We now consider the problem of optimizing η_{in} for the desired fully complex spot-array generator. We first note that a broad class of input transmittances can be written as

$$g(x, \phi) = \frac{f(x, \phi)}{\max_x|f(x, \phi)|}, \qquad (14)$$

where $g(x, \phi)$ is the generating or target transmittance, $f(x, \phi)$ is a generating function appropriate for the designated task, and ϕ is a free, real, vector parameter. We also note that, regardless of the value of ϕ, $|g(x, \phi)| \leq 1$, as befitting a passive device. Under some circumstances ϕ can be used to optimize the performance of the device. For a spot-array generator that furnishes an array of far-field spots at, say, $\{u_i,\ i = 1,\ldots, M\}$, an appropriate generating function is

$$f(x, \phi) = \sum_{k=1}^{M} \exp[j(2\pi u_k x + \phi_k)], \qquad (15)$$

where $\phi = (\phi_1,\ldots, \phi_k,\ldots, \phi_M)^{\mathrm{T}}$. For $\phi = \mathbf{0}$, i.e., each component of ϕ is zero, the η_{in} computation yields approximately $\eta_{\mathrm{in}} = 1/M$. A significant improvement can be obtained by a judicious choice of ϕ. Indeed, inserting

Eq. (15) in Eq. (14) and using Eq. (1) yield an input diffraction efficiency described approximately by

$$\eta_{\text{in}}(\phi) = \alpha^2(\phi)M, \tag{16}$$

where

$$\alpha(\phi) \equiv \frac{1}{\max_x |f(x, \phi)|}. \tag{17}$$

Thus, to maximize η_{in}, we seek to find ϕ^* such that

$$\phi^* = \arg[\max_\phi \alpha^2(\phi)]. \tag{18}$$

Equation (18), in words, says that ϕ^* is the vector that will make $\alpha^2(\phi)$ as large as possible. Clearly, from Eq. (16), this will maximize η_{in}. However, an analytic solution to Eq. (18) is not readily apparent. Thus, in the rest of this paper, we consider approximate solutions to Eq. (18). In particular, we focus on the three methods described below: Monte Carlo simulation, genetic algorithms, and gradient descent.

A. Monte Carlo Simulation

In the Monte Carlo simulation a random-number (RN) generator is used to generate a new random phase vector at each trial. This phase vector is then used to evaluate η_{in}. The largest η_{in} and the corresponding phase vector are retained. Two experiments were performed: one of 1000 trials and one of 10,000 trials. The phases and the corresponding diffraction efficiencies are recorded for both experiments. The results are given in Section 6.

B. Genetic Algorithm

Details of the genetic algorithm (GA) used for optimizing a 10-spot array are given in Appendix B. Here we furnish only a summary of the parameters used: initial population size is 25, each member being a 10-component vector of phases, each represented by a 10-bit binary string; fitness scaling parameter is 2; fitness function is η_{in}; initial mutation probability is 0.01; and final mutation probability is 0.005.

C. Gradient Method

A gradient algorithm requires an objective function that we seek to extremize. Such an objective function can be constructed by the following reasoning. Recall Eq. (2) for η_{in}; we wish to make η_{in} as large as possible. Now $\eta_{\text{in}} = 1$ if $|g| = \alpha|f| = 1$ for all x or, equivalently, $|f|^2 = 1/\alpha^2 \equiv \beta$. Therefore, more generally, $\beta - |f|^2$ is the error from the optimum at location x, and

$$e(\beta, \phi) \equiv \sum_{k=1}^{N} [\beta - |f(k\Delta, \phi)|^2]^2 \tag{19}$$

is proportional to the total mean square error over all x. Thus $e(\beta, \phi) > 0$ is a suitable objective function that has to be minimized over β and ϕ. To find β and ϕ that would minimize the objective function, we use the iterative gradient formula:

$$(\beta, \phi)_{k+1} = (\beta, \phi)_k - \gamma \nabla e(\beta, \phi), \tag{20}$$

where

$$\nabla e(\beta, \phi) = \left(\frac{\partial e}{\partial \beta}\, \frac{\partial e}{\partial \phi_1} \cdots \frac{\partial e}{\partial \phi_M}\right)^{\mathrm{T}},$$

k is the iteration number, γ is a constant step size, and the gradient $\nabla \epsilon(\beta, \phi)$ is evaluated at $(\beta, \phi)_k$. The initial starting vector $(\beta, \phi)_0$ is chosen by using a RN generator to supply values for the components.

5. COMPUTATIONAL EFFICIENCY: COMPARISON

It is of interest to consider how the different optimization routines compare vis-à-vis the amount of computation. In this section we furnish an analysis of the relative computational efforts involved in each of the three optimization algorithms for a 10-spot array.

A. Monte Carlo Search

Let T_ϕ denote the time that it takes to draw 10 random phases and $T_{\max}$ the time needed to find the maximum value of the generating function $f(x, \phi)$. For each phase vector ϕ, we need to compute the diffraction efficiency in Eq. (2). Let $T_{\eta_{\text{in}}}$ denote the time to compute η_{in}. Then the approximate computation time per cycle is

$$T_\phi + T_{\max} + T_{\eta_{\text{in}}},$$

and for C cycles the total time would be

$$T_T^{(\text{MC})} = C(T_\phi + T_{\max} + T_{\eta_{\text{in}}}). \tag{21}$$

In our simulation we tried two values of C: 1000 and 10,000.

B. Genetic Algorithm

For each of the 25 strings, we must draw a 10-component random phase vector and evaluate the associated η_{in}. There are 25 probability computations and 25 scaling operations. Then with T_s, T_{co}, and T_{mu} denoting the time for scaling, crossover, and mutations, the total time for the GA is

$$\begin{aligned} T_T^{(\text{GA})} &= 25Q(T_\phi + T_{\max} + T_{\eta_{\text{in}}} + T_s + T_{\text{co}} + T_{\text{mu}}) \\ &\cong 25Q(T_\phi + T_{\max} + T_{\eta_{\text{in}}}), \end{aligned} \tag{22}$$

where Q is the number of generations required to achieve convergence. We have assumed that $T_s + T_{\text{co}} + T_{\text{mu}} \ll T_\phi + T_{\max} + T_{\eta_{\text{in}}}$. This is borne out by our simulations. The value of Q typically varied near 80.

C. Gradient Search

This case is more difficult to analyze, since the computational load is related to the complexity of the derivative of the generating function. Fortunately, for the 10-spot array, the derivative of the generating function has an analytic form quite similar to that of the generating function itself. Indeed, for the 10-spot array, we can derive that

$$
\begin{aligned}
\eta_{\mathrm{in}}(\phi) &= \sum_{k=1}^{N} \alpha^2 |f(k\Delta, \phi)|^2 \\
&= \sum_{k=1}^{N} \alpha^2 \sum_{l=1}^{M} \sum_{i=1}^{M} \\
&\quad \exp\{j[2\pi(u_l - u_i)k\Delta + \phi_l - \phi_i]\} \\
&= \sum_{k=1}^{N} \alpha^2 [M + 2Q(k\Delta)], \qquad (23)
\end{aligned}
$$

where $Q(k\Delta)$ is given by

$$
Q(k\Delta) = \sum_{l=1}^{M} \sum_{i=1+1}^{M} \cos[2\pi(u_l - u_i)k\Delta + \phi_l - \phi_i]. \qquad (24)
$$

Likewise, we can write the following for the objective function for the 10-spot array:

$$
e(\beta, \phi) = \sum_{k=1}^{N} [\beta - M - 2Q(k\Delta)]^2, \qquad (25)
$$

$$
\frac{\partial e}{\partial \beta} = \sum_{k=1}^{N} [2\beta - 2M - 4Q(k\Delta)], \qquad (26)
$$

$$
\frac{\partial e}{\partial \phi_l} = \sum_{k=1}^{N} [2\beta - 2M - 4Q(k\Delta)][2S_l(k\Delta)], \qquad l = 1,\ldots, M, \quad (27)
$$

where

$$
S_l(k\Delta) = \sum_{\substack{i=1 \\ i \neq l}}^{M} \sin[2\pi(u_l - u_i)k\Delta + \phi_l - \phi_i], \qquad l = 1,\ldots, M. \quad (28)
$$

Comparing Eqs. (27) and (28) with Eq. (23), we see that computation of each gradient component is roughly on the order of $T_{\eta_{\mathrm{in}}}$. The total number of derivative computations is $M + 1$, but since one component of ϕ can be set to zero relative to the others, the actual number is M and the computation time per cycle is $MT_{\eta_{\mathrm{in}}}$. If there are P cycles required for convergence, the total time required is

$$
T_T^{(\mathrm{GR})} = PMT_{\eta_{\mathrm{in}}}. \qquad (29)
$$

In our computations P had the value $P = 20$ to 50.

6. OUTPUT DIFFRACTION EFFICIENCY

While the input diffraction efficiency η_{in} is computed as the energy of the target function, it is useful to have another measure of efficiency that is defined strictly in terms of diffracted light. Such a measure is furnished by the output diffraction efficiency η_{out}, defined by

$$
\eta_{\mathrm{out}} = \frac{\text{energy in desired diffraction pattern}}{\text{total energy in frequency plane}} = \frac{\int_R I(u)\mathrm{d}u}{\int_{-\infty}^{\infty} I(u)\mathrm{d}u}, \qquad (30)
$$

where R is the region containing the desired diffraction pattern. We would like to make η_{out} independent of aperture size, thereby having it reflect the properties of only the generating transmittance $g(x)$. One convenient way to do this is to make the input aperture infinitely large. Thus, in the case of a diffraction pattern consisting of spot arrays, the input aperture would contain an infinitely periodic spatial function whose precise character would depend on the energy distribution among the spots.

One final remark is in order before proceeding. The two efficiencies η_{in} and η_{out} are closely related. This is implied by Eq. (7), which, while it is a result based on an average, shows that as η_{in} goes to unity, essentially all of the diffracted light consists of the prescribed portion $[\overline{G_1(u)}]^2$ as opposed to noise.

7. NUMERICAL RESULTS

Equation (15), with $M = 10$, is used as the generating function for a 10-spot array. The three algorithms—Monte Carlo simulation, genetic, and gradient search—were implemented by using the following parameters: aperture size, $L = 1$; pixel size in the SLM plane, $\Delta x = 0.00195$; pixel size in the Fourier plane, $\Delta u = 1$; number of points N in the discrete Fourier transform, $N = 512$; spot locations in the frequency plane, $u_1 = 210$, $u_2 = 220$, $u_3 = 230$, $u_4 = 240$, $u_5 = 250$, $u_6 = 260$, $u_7 = 270$, $u_8 = 280$, $u_9 = 290$, and $u_{10} = 300$; and number of points in a lookup table of $\sin c^{-1}(x)$, 5000, uniformly spread over the range (0,1). With these data and the material in Table 1, it is possible to replicate all the results in Table 2 and the figures.

Table 1 gives the phase vectors computed for each algorithm. Table 2 gives the performance results for the

Table 1. Optimum Phase Angles[a]

Algorithm	ϕ_1	ϕ_2	ϕ_3	ϕ_4	ϕ_5	ϕ_6	ϕ_7	ϕ_8	ϕ_9	ϕ_{10}
MCS1	0.9729	0.5579	0.7178	0.1777	0.5698	0.3401	0.2559	0.5098	0.5428	0.6863
MCS2	0.3224	0.8596	0.6905	0.2020	0.3622	0.7764	0.3858	0.4821	0.3755	0.3600
GA	0.7243	0.6784	0.6149	0.7331	0.1056	0.4018	0.7185	0.2845	0.9130	0.4770
Gradient	0.6705	0.2675	0.3119	0.8618	0.1963	−0.1426	0.6288	0.8404	−0.0690	0

[a] Final phases obtained by three algorithms: MCS1 is the Monte Carlo simulation with 1000 tries, MCS2 is the Monte Carlo simulation with 10,000 tries, GA is the genetic algorithm, and Gradient is the gradient-search algorithm used with the objective function in Eq. (19). To get the actual phase in radians, multiply each entry by 2π.

Table 2. Performance of Optimization Routines[a]

Performance Measure	Algorithm				
	MCS1	MCS2	GA	Gradient	NPRE
Pseudorandom Encoding					
η_{in} [Eq. (2)]	47.67%	57.98%	64.65%	69.33%	10%
Average η_{out}	47.85%	58.57%	64.96%	69.35%	10.65%
$\overline{\text{ADC}}$ [Eq. (33)]	0.2251	0.1897	0.1597	0.1846	0.5427
ADC* [Eq. (34)] (worst case)	0.3270	0.2924	0.2685	0.2497	0.7942
$\overline{\sigma_N}$ [Eq. (35)]	0.1363	0.1144	0.1010	0.1151	0.3479
Kinoform (Phase-Only)					
η_{out} [Eq. (31)]	94.17%	95.61%	97.96%	96.56%	53.94%
ADC [Eq. (32)]	0.2948	0.3805	0.4616	0.2135	0.9179
σ_N [Eq. (36)]	0.2112	0.2223	0.3013	0.1171	1.3562
Speed	1000	10000	~2000	200–500	0

[a] Input (η_{in}) and output (η_{out}) diffraction efficiencies for the four optimization routines as well as for NPRE and phase-only approaches. The *actual* speed obviously depends on many factors, such as the platform used, the software, and the skill of the programmer. The numbers in the Speed row are the number of cycle times ($T_\phi + T_{\max} + T_{\eta_{\text{in}}}$) required to achieve convergence. For the gradient case the cycle time is actually shorter, since $T_\phi + T_{\max}$ is absent.

three optimization algorithms, as well as for the direct PO (kinoform) result and naive pseudorandom encoding (NPRE). However, before we discuss the results, some remarks are in order:

1. Neither η_{in} nor η_{out} measures the uniformity of the spot intensities in the spot array. Indeed, a large η_{out} implies that most of the light is going to the correct locations but is not indicative that the light is evenly distributed. To control the uniformity of the peaks, one should define a diffraction-plane uniformity metric and use this metric as a constraint in the aperture plane. This is what is done in an iterative vector-space algorithm. We shall call this metric the average deviation contrast (ADC) and define it below.
2. Under ideal circumstances one would expect that, by Parseval's theorem, η_{in} and η_{out} should be essentially equivalent. However, the presence of PRE noise will usually cause a minor difference between them. Since noise intensity adds to the total power in the peaks, it is not surprising to find η_{out} to be slightly larger than η_{in}.

Table 2 needs some explanation. The first row lists the four optimization routines: MCS1 (Monte Carlo simulation with 1000 trials), MCS2 (Monte Carlo simulation with 10,000 trials); GA (genetic algorithm), the gradient search, and, as reference, NPRE, i.e., set $\phi = 0$. The third row lists η_{in} for various cases, with the use of Eq. (2) for η_{in} and Eqs. (14) and (15) for the generating transmittance and the generating function, respectively. The fourth row yields η_{out} for the various cases, with the use of a discrete equivalent of Eq. (30) for the spot array:

$$\eta_{\text{out}} = \frac{\sum_{i=1}^{10} I(u_i)}{\sum_{n=0}^{512} I(n\Delta u)}. \tag{31}$$

Since each PRE trial yields a random outcome, the η_{out} that appears in the fourth row is the average of 10 such trials.

The next three rows are measures of the uniformity of the spots in the spot array. For example, row 5 yields the ADC averaged over 10 trials. The ADC for a single trial, say the ith, is computed as

$$\text{ADC}_i = \frac{I^{(i)}_{\max,p} - I^{(i)}_{\min,p}}{I^{(i)}_{\max,p} + I^{(i)}_{\min,p}}, \tag{32}$$

where $I^{(i)}_{\max,p}$ and $I^{(i)}_{\min,p}$ are the intensities of the highest and lowest peaks at the desired locations, respectively, of the ith trial. A low value of ADC is desired (zero is optimum); a high value indicates considerable variability among peaks. Then row 5 gives

$$\overline{\text{ADC}} = \frac{1}{10}\sum_{i=1}^{10} \text{ADC}_i, \tag{33}$$

while row 6 gives the worst case,

$$\text{ADC*} = \max_i \text{ADC}_i. \tag{34}$$

The entries in row 7 yield

$$\overline{\sigma_N} = \frac{1}{10}\sum_{i=1}^{10} \sigma_N^{(i)}, \tag{35}$$

where $\sigma_N^{(i)}$, the normalized standard deviation, is given by

$$\sigma_N^{(i)} = \frac{\left\{\sum_{n=1}^{10} [I^{(i)}_{\text{peak},n} - \overline{I}^{(i)}_{\text{peak}}]^2\right\}^{1/2}}{\overline{I}^{(i)}_{\text{peak}}}, \tag{36}$$

where $I^{(i)}_{\text{peak},n}$ is the intensity of the nth peak (there are 10) of the ith trial (there are also 10) and $\overline{I}^{(i)}_{\text{peak}}$ is the average peak height of the ith trial, computed as

$$\overline{I}^{(i)}_{\text{peak}} = \frac{1}{10}\sum_{n=1}^{10} I^{(i)}_{\text{peak},n}. \tag{37}$$

In Eq. (36) n is a spatial, transverse index, while i is a longitudinal time or ensemble index. Also, in Eqs. (33)–(35), small numbers are more desirable than large ones.

Based on an examination of the first seven rows of Table 2, it appears that all four optimization routines greatly outperform NPRE. In attempting to evaluate the results, however, it is important to remember that not only is PRE a random process but all the optimization routines contain an element of randomness as well. MCS1, MCS2, and the GA are inherently stochastic procedures, and even the gradient search requires a random starting point. To illustrate the effect of the random starting point on the gradient-search results, we obtained $\eta_{\text{in}} = 68.86\%$ after 20 iterations in one run, $\eta_{\text{in}} = 67.74\%$ after 50 iterations in another run, and $\eta_{\text{in}} = 68.33\%$ in a third run. The values of η_{in} in the third row are the *best* results selected from several trials.

It would appear that the gradient search and the GA clearly outperform MCS1 and MCS2. While the gradient search yields a slightly higher η_{in} than the GA (69% versus 65%), the uniformity of the peaks is slightly better in the GA (ADC = 0.16 versus 0.18 for the gradient algorithm; remember that smaller is better).

The three rows under the heading Kinoform (Phase-Only) yield performance data on the PO case. In other words, having computed the phases by one of the various optimization methods, we realize an input transmittance as $\exp\{j \arg[g(x, \phi)]\}$. Thus the input transmittance is modulated by only the *phase* of the generating function. The most striking observation here is the very high η_{out}. Indeed, it is not difficult to show that, for the PO method, high values of η_{out} are expected.[7] The downside of the PO approach is the possibility of significant unevenness of the peaks in the spot array as compared with the peaks produced by the optimized PRE. The advantage of the latter over the former becomes evident upon comparing rows 4 and 6 with rows 8 and 9, respectively.

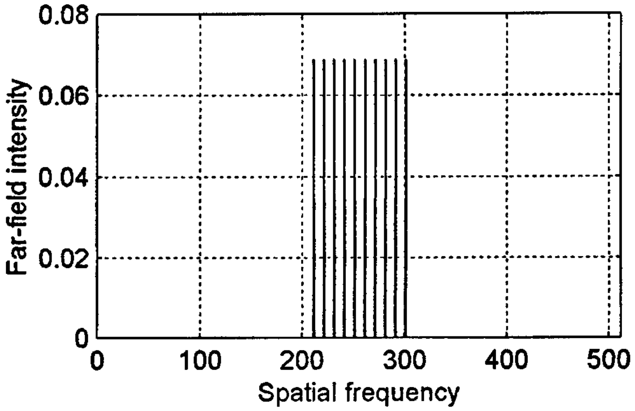


Fig. 1. Diffraction intensity produced by an ideal generating function [Eq. (25)] with no phase optimization.

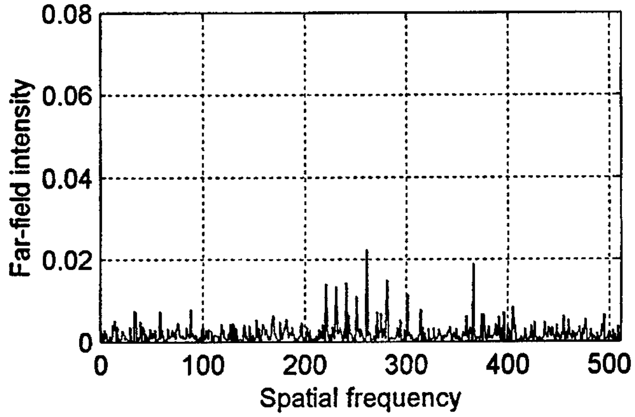


Fig. 2. Diffraction intensity produced by the NPRE algorithm.

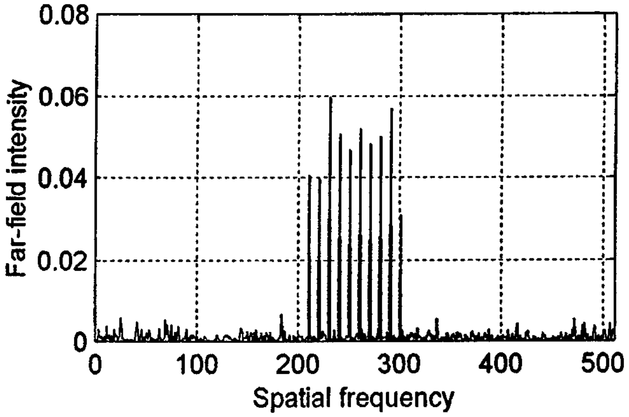


Fig. 3. Diffraction intensity realized by PRE after phase optimization by MCS1.

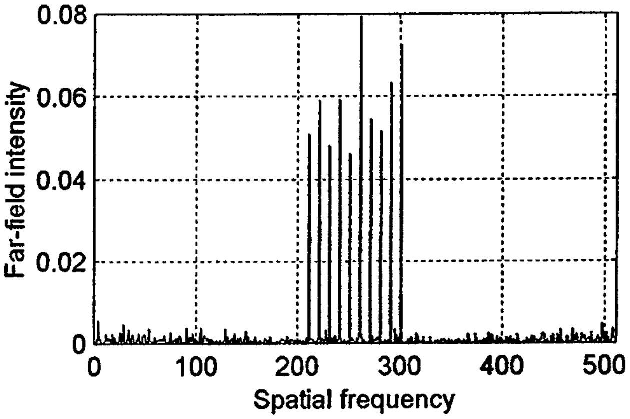


Fig. 4. Diffraction intensity realized by PRE after phase optimization by MCS2.

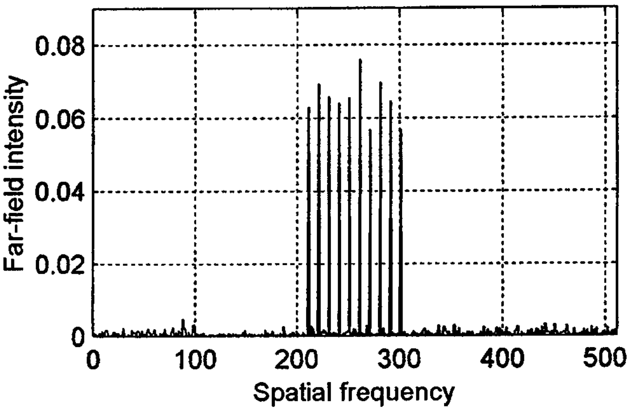


Fig. 5. Diffraction intensity after phase optimization by the genetic algorithm followed by PRE.

Graphical results reinforce the conclusions presented in Table 2. Figure 1 shows the spot array generated by the generating transmittance of Eq. (14) with the use of Eq. (15). In this figure the phases are those computed by the gradient-search algorithm (Table 1). As expected, the peaks are uniform and the cross-term noise (overlapping sidelobes) is virtually imperceptible on a linear scale. Figure 2 shows the "spot array" generated by the NPRE algorithm: The spot array is difficult to detect in the noise; hence the need to optimize the generating function. Figures 3 and 4 show the spot-array peaks with the use of PRE after optimizing by MCS1 and MCS2, respectively. Here the values of η_{out} are 47.71% and 58.85%, respectively. The unevenness of the peaks is greatly reduced over that of the NPRE case but is still pronounced when MCS1 is used. The efficiency in using MCS2 is noticeably higher than that in using MCS1.

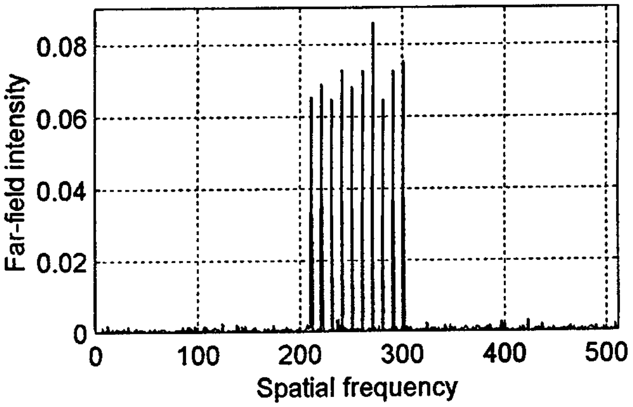


Fig. 6. Diffraction intensity after phase optimization by the gradient algorithm followed by PRE.

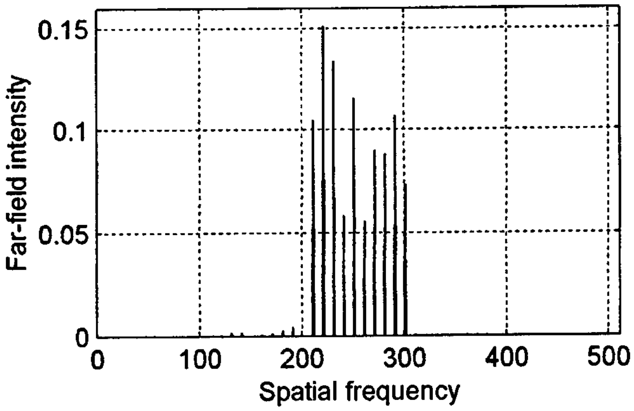


Fig. 7. Phase-only diffraction intensity produced by a genetic-algorithm-optimized generating transmittance.

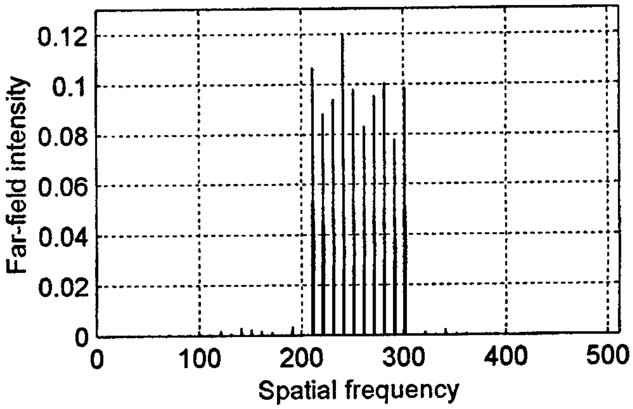


Fig. 8. Phase-only diffraction intensity produced by a gradient-search-optimized generating transmittance.

Figure 5 shows the diffraction intensity furnished by PRE with use of the GA-optimized generating transmittance for $\eta_{\text{out}} = 65.42\%$ and $\bar{\sigma} = 0.0834$. Figure 6 shows the PRE result with use of the gradient-optimized generating transmittance. It has $\eta_{\text{out}} = 71.48\%$ and $\bar{\sigma} = 0.0848$. It can be seen from Fig. 5 that the GA yields the most uniform spot array, while from Fig. 6 we see that the gradient search yields the highest diffraction efficiency.

Finally, Figs. 7 and 8 show the direct PO results with use of the optimized phases from the GA and the gradient search, respectively. As stated above, the diffraction efficiency η_{out} is very high in both cases, but the nonuniformity of the peaks is pronounced in the PO/GA case and might be unacceptable for optical switching or related applications. Interestingly, the PO/gradient-search case gave acceptable results.

8. CONCLUSIONS

Pseudorandom encoding (PRE) is a means for approximately realizing a desired far-field diffraction pattern by modulating only the phase of the input transparency. As a consequence, the prescribed far field is realized, on the average, but in the presence of noise. Optimizing the input diffraction efficiency, the latter being proportional to the energy in the generating transmittance, can reduce the noise and increase the diffraction efficiency. In the case of a spot array, the generating or target function contains an adjustable free-phase vector whose proper selection can lead to higher diffraction efficiencies. Unfortunately, a proper selection by analytic means does not seem possible.

In this paper we considered several techniques for finding the optimum free-phase vector. Best results were obtained by using both a genetic algorithm and a gradient search, which can offer significant improvements over naive PRE and even Monte Carlo simulation. Of the three optimization methods studied, the gradient-search algorithm has the lowest computational complexity.

APPENDIX A: REVIEW OF THE PSEUDORANDOM ENCODING ALGORITHM

There are two key ideas behind PRE: (1) that the expected value of a discrete random variable can be different from any of the values that the random variable can realize and (2) the law of large numbers. How the PRE is affected by the LLN is discussed in Section 3 of the paper. To illustrate the first idea, however, is easy. Suppose that we wish to realize the transmittance value $g(x) = 0.745 \exp(i0.46)$ at a point x but we are limited to a unity-magnitude transmittance and a phase that can take only one of two values: 0 and $\pi/2$. Then with $p = \text{Prob}(\theta = \pi/2)$ and $q = 1 - p = \text{Prob}(\theta = 0)$, we find that $E[\exp(i\theta)] = 0.745 \exp(i0.46)$ when $p = 1/3$. Thus the transmittance at x has the required value as an *ensemble average* but has a value of either 1 or $j = \sqrt{-1}$ for *any realization.* The mean square error ϵ_{ms} for this example has value $\epsilon_{\text{ms}} = 0.444$. In general, except for some trivial cases, there will always be an error when a fully complex function is realized, on the average, by using an ensemble of PO values.

In the studies of Cohn and co-workers,[23–25] PRE is taken to mean the procedure by which the phases of uniform-magnitude SLM-plane pixels are chosen from an appropriate uniform distribution to achieve an average far-field intensity that approximately corresponds to the prescribed far field.

Realizations of the desired fully complex transmittance function $g(x)$ are achieved through random-sample functions $\{\exp[j\theta(x)]\}$. In particular, at each x, we seek to solve the equation

$$g = \int_{-\infty}^{\infty} \exp(j\alpha) f_\theta(\alpha) \mathrm{d}\alpha \tag{A1}$$

for the probability density function $f_\theta(\alpha)$, where $g = g(x)$, $\theta = \theta(x)$, etc. A solution to Eq. (A1) is obtained by limiting the solution to the subset of the two-parameter (c, w) family of uniform probability density functions of the form

$$\phi_\theta(\alpha; c, w) = \frac{1}{w} \mathrm{rect}\left(\frac{\alpha - c}{w}\right), \tag{A2}$$

where c is the mean and $w^2/12$ is the variance of the associated random variable θ. Using Eq. (A2) in Eq. (A1) yields

$$g = \exp(jc) \frac{\sin(w/2)}{w/2} = \exp(jc)\mathrm{sinc}(w/2\pi). \tag{A3}$$

Then $\exp(j\theta)$ is an unbiased estimator for g if, at the point x, the value of θ is chosen from a uniform probability density function with parameters c and w such that

$$c(x): \quad c = \arg[g(x)], \tag{A4}$$

$$w(x): \quad \mathrm{sinc}(w/2\pi)| = |g(x)|. \tag{A5}$$

In particular, with (c, w) restricted to $0 \leqslant c < 2\pi$ and $0 \leqslant w < 2\pi$ for all x in the support of $g(x)$, the equation $|g(x)| = \mathrm{sinc}(w/2\pi)$ is invertible as $w = 2\pi\, \mathrm{sinc}^{-1}(|g|)$ for $|g(x)| \leqslant 1$.

APPENDIX B: DESCRIPTION OF THE GENETIC ALGORITHM

The steps in implementing the GA for a 10-spot array involve the following:

1. With $\phi_i = (\phi_{i1}, \phi_{i2}, \dots, \phi_{i10})$ representing the ith trial phase vector, each component ϕ_{ij}, $j = 1, \dots, 10$, is assigned a 10-bit binary string allowing for the representation of 1024 possible phase values. This is repeated for $i = 1, 2, \dots, 25$. The choice of 25 population elements is somewhat arbitrary. The actual values of the phases are obtained from a RN generator and converted to binary form. The totality of binary characters representing the phase vector ϕ_i is called a string. The 25 strings, representing the 25 phase vectors ϕ_i, $i = 1, \dots, 25$, form an initial population.

2. For each ϕ_i, selection probabilities $\{p_i\}$ and cumulative selection probabilities $\{q_i\}$ are computed, respectively, as

$$p_i = \frac{\eta_{\text{in}}^{(i)}}{\sum\limits_{i=1}^{25} \eta_{\text{in}}^{(i)}}, \qquad q_k = \sum_{i=1}^{k} p_i, \tag{B1}$$

where $\eta_{\text{in}}^{(i)}$ is the fitness function of the ith trial phase vector.

3. A new set of fitness values and p_i^*'s and q_k^*'s are created with the use of a linear function to transform the previous set $\{\eta_{\text{in}}^{(i)}\}$ into a new set $\{\eta_{\text{in}}^{(i)*}\}$. The linear function is given by

$$\eta_{\text{in}}^{(i)*} = \alpha \eta_{\text{in}}^{(i)} + \beta,$$

where

$$\alpha = \frac{C\eta_{\max} - \eta_{\text{avg}}}{\eta_{\max} - \eta_{\text{avg}}},$$

$$\eta_{\max} = \max_i \eta_{\text{in}}^{(i)},$$

$$\eta_{\text{avg}} = \frac{1}{25} \sum_{i=1}^{25} \eta_{\text{in}}^{(i)},$$

$$\beta = 1 - \alpha, \tag{B2}$$

and C is a constant determined by the user. Typically, for small populations, C is adjusted to lie in the interval 1.2–2. In our case $C = 2$. The purpose of fitness scaling is to avoid premature convergence as well as maintaining a significant number of high-fitness strings late in the run.

4. With the use of a RN generator to generate numbers in the interval [0,1], a set of 25 strings is selected from the original population according to the newly created cumulative selection probabilities $\{q_k^*\}$.

5. Crossover: Offspring are created from mating pairs randomly selected from the 25 strings. Crossover sites are determined by using outputs of a RN generator issuing integers in the range [1,99].

6. Mutations: These are usually created with very low probabilities. In our case the mutation probability was 0.01 and reduced to 0.005 as the procedure matured.

7. Repeat steps 2–6 to generate subsequent generations. Stop when the convergence criterion has been met.

ACKNOWLEDGMENTS

The work in this paper was supported by Office of Naval Research grant N00014-96-1-1296, a Defense Advanced Research Projects Agency grant through ARO DAAH04-94-G-0358, and NASA cooperative agreement NCC5-222. The authors are grateful to three anonymous reviewers for their perceptive comments, which led to a better paper.

Address correspondence to Yongyi Yang at the location on the title page or by phone, 312-567-3423 or e-mail, yy@ece.iit.edu.

REFERENCES

1. J. N. Mait, "Understanding diffractive optic design in the scalar domain," J. Opt. Soc. Am. A **12**, 2145–2158 (1995).
2. J. N. Mait, "Fourier array generators," in *Micro-Optics: Elements, Systems, and Applications*, H. P. Herzig, ed. (Taylor & Francis, London, 1997), pp. 293–323.
3. U. Krackhardt, J. N. Mait, and N. Streibl, "Upper bound on the diffraction efficiency of phase-only fanout elements," Appl. Opt. **31**, 27–37 (1992).
4. F. Wyrowski, "Diffraction efficiency of analog and quantized digital amplitude holograms: analysis and manipulation," J. Opt. Soc. Am. A **7**, 383–393 (1990).
5. F. Wyrowski and O. Bryngdahl, "Iterative Fourier-transform algorithm applied to computer holography," J. Opt. Soc. Am. A **5**, 1058–1065 (1988).
6. F. Wyrowski, "Iterative quantization of digital amplitude holograms," Appl. Opt. **28**, 3864–3870 (1989).
7. F. Wyrowski, "Upper bound of the diffraction efficiency of diffractive phase elements," Opt. Lett. **16**, 1915–1917 (1991).
8. J. P. Allebach and D. W. Sweeney, "Iterative approaches to computer generated holography," in *Computer-Generated Holography II*, S. H. Lee, ed., Proc. SPIE **884**, 2–9 (1988).
9. H. Stark, W. C. Catino, and J. L. LoCicero, "Design of phase gratings by generalized projections," J. Opt. Soc. Am. A **8**, 566–571 (1991).
10. H. Stark, Y. Yang, and D. Gurkan, "Factors affect convergence in the design of diffractive optics by iterative vector space methods," J. Opt. Soc. Am. A **16**, 149–159 (1999).
11. N. C. Gallagher and B. Liu, "Method for computing kinoforms that reduces image reconstruction error," Appl. Opt. **12**, 2328–2335 (1973).
12. M. P. Dames, R. J. Dowling, P. McKee, and D. Wood, "Efficient optical elements to generate intensity weighted spot arrays: design and fabrication," Appl. Opt. **30**, 2685–2691 (1991).
13. J. Bengtsson, "Kinoform design with an optimal-rotation-angle method," Appl. Opt. **33**, 6879–6884 (1994).
14. B. R. Brown and A. W. Lohmann, "Complex spatial filtering with binary masks," Appl. Opt. **5**, 967–970 (1966).
15. W. H. Lee, "Computer-generated holograms: techniques and applications," in *Progress in Optics*, E. Wolf, ed. (Elsevier, Amsterdam, 1978), Vol. 16, pp. 119–231.
16. W. J. Dallas, "Computer-generated holograms," in *The Computer in Optical Research*, B. R. Frieden, ed. (Springer, Berlin, 1980), Chap. 6, pp. 291–366.
17. L. B. Lesem, P. M. Hirsch, and J. A. Jordon, Jr., "The kinoform: a new wavefront reconstruction device," IBM J. Res. Dev. **13**, 150–155 (1969).
18. D. C. Chu and J. R. Fienup, "Recent approaches to computer-generated holograms," Opt. Eng. **13**, 189–195 (1974).
19. D. Casasent and W. A. Rozzi, "Computer-generated and phase-only synthetic discriminant function filters," Appl. Opt. **25**, 3767–3772 (1986).
20. D. Prongue, H. P. Herzig, R. Dandliker, and M. T. Gale, "Optimized kinoform structures for highly efficient fan-out elements," Appl. Opt. **31**, 5707–5711 (1992).
21. M. W. Farn, "New iterative algorithm for the design of phase-only gratings," in *Holographic Optics: Computer and Optically Generated*, I. Cindrich and S. H. Lee, eds., Proc. SPIE **1555**, 34–42 (1991).
22. J. D. Stack and M. R. Feldman, "Recursive mean-squared-error algorithm for iterative discrete on-axis encoded holograms," Appl. Opt. **31**, 4839–4846 (1992).
23. R. W. Cohn and L. G. Hassebrook, "Representations of fully complex functions on real-time spatial light modulators," in *Optical Information Processing*, F. T. S. Yu and S. Jutamulia, eds. (Cambridge U. Press, Cambridge, UK, 1998), Chap. 15, pp. 396–432.
24. R. W. Cohn and M. Liang, "Approximating fully complex spatial modulation with pseudo-random phase-only modulation," Appl. Opt. **33**, 4406–4415 (1994).
25. R. W. Cohn and M. Liang, "Pseudo-random phase-only encoding of real-time spatial light modulators," Appl. Opt. **35**, 2488–2498 (1996).
26. S. Geman and D. Geman, "Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images," IEEE Trans. Pattern. Anal. Mach. Intell. **PAMI-6**, 721–741 (1984).
27. S. Kirpatrick, C. D. Gelatt, and M. P. Vecchi, "Optimization by simulated annealing," Science **220**, 671–680 (1983).
28. H. Stark and Y. Yang, *Vector Space Projections: A Numerical Approach to Signal and Image Processing, Neural Nets, and Optics* (Wiley, New York, 1998).

Modified minimum-distance criterion for blended random and nonrandom encoding

Markus Duelli, Matthew Reece, and Robert W. Cohn

The ElectroOptics Research Institute, University of Louisville, Louisville, Kentucky 40292

Received December 7, 1998; accepted May 17, 1999; revised manuscript received May 24, 1999

Two pixel-oriented methods for designing Fourier transform holograms—pseudorandom encoding and minimum-distance encoding—usually produce higher-fidelity reconstructions when combined than those produced by each method individually. In previous studies minimum-distance encoding was defined as the mapping from the desired complex value to the closest value produced by the modulator. This method is compared with a new minimum-distance criterion in which the desired complex value is mapped to the closest value that can be realized by pseudorandom encoding. Simulations and experimental measurements using quantized phase and amplitude modulators show that the modified approach to blended encoding produces more faithful reconstructions than those of the previous method. © 1999 Optical Society of America

[S0740-3232(99)02010-4]

OCIS codes: 230.6120, 090.1760, 030.6600, 070.2580.

1. INTRODUCTION

Today the leading methods of designing Fourier transform holograms for laser pattern generation and optical interconnects use iterative search and numerical optimization procedures that vary the modulation values and various degrees of freedom to achieve acceptable diffraction patterns.[1–6] In this prior work it is normally assumed that the design is to be realized as a fixed-pattern diffractive optical element that is subsequently mass produced, which makes computation times of a few minutes to hours[7] insignificant compared with the time required to fabricate the device. However, our previous studies on real-time programmable spatial light modulators (SLM's)[8] and diffractive optical element rapid prototyping systems[9] have led us to reconsider the design problem with particular emphasis on significantly reducing the design time.

By far, the fastest design algorithms are those that directly map a desired complex-valued function into a transmittance function that can be physically produced by the available modulator. The delayed-sampling method of Brown and Lohmann is one of the earliest applications in optics of this idea.[10] The numerical speed of this and many other mapping/encoding methods that were evaluated in the first decade of computer-generated holography[11,12] is due to serial encoding of each desired complex value into a corresponding value of transmittance. Since the various degrees of freedom are not included in this design approach (e.g., in the design of most spot array generators, where the phase of the far-field diffraction pattern is usually not of concern), the performance of the encoding method in terms of diffraction efficiency or other related metrics can be substantially less than that for the optimization methods. Nonetheless, we believe that there are applications that would benefit from the faster encoding algorithms (for example scenarios see Ref. 13).

To reduce the differences in performance between optimal designs and encoded designs, we have begun investigating suboptimal design strategies in which some additional computations (but fewer than those required to find the global optimum) are directed at improving device performance.[14,15] Two possible suboptimal strategies are (1) to use the best solution found by optimization for a given amount of time or (2) to optimize by using some, rather than all, of the available design freedoms. It is this second approach that we consider here.

Specifically, we consider encoding methods that can be improved by varying a single design freedom/free parameter. The free parameter (referred to as γ) scales the magnitude of the complex-valued function that is to be encoded. Each value of the complex function is encoded by one of two encoding algorithms: pseudorandom encoding (PRE)[16] for smaller-magnitude values and minimum-distance encoding (MDE)[17] for larger values, both of which will be reviewed in Sections 2 and 3. Increasing the value of the free parameter decreases the number of complex values that are encoded by the PRE algorithm and increases the number of values encoded by the MDE algorithm. In this way the free parameter controls the blending of the two encoding algorithms.

Designs by this approach have been described in a nonarchival conference proceedings[14] and in brief detail in a short paper.[15] In each specific design considered, it was found that better performance is achieved over PRE or MDE individually by blending the two algorithms and that there is a particular degree of blending (as measured by γ) that gives the best performance of all blendings. At times we have noted dramatic improvements in performance even if only a few percent of the complex values are encoded by MDE.[14] However, we recently observed for modulators that produce only three quantized values of phase that the blending of MDE and PRE leads to only slight performance improvements, and for some blendings the performance is even lower than that without blending.

0740-3232/99/102425-14$15.00

This observation leads us in this paper to propose, consider, and evaluate a modified blending of encoding algorithms. Figure 1(a) shows the phase-only SLM characteristic that was considered in Refs. 14 and 15. Desired complex values that are inside the phase-only modulation characteristic (striped region) can be pseudorandom encoded. The values outside the region are mapped to the closest point on the modulation characteristic (along radial lines centered on the origin).

There is an alternative possible mapping that becomes apparent when considering blended encoding with noncircular SLM characteristics. This is illustrated in Fig. 1(b) for a tri-phase SLM. The striped region again represents the range of values that can be encoded by the PRE algorithm. There are now two possible minimum-distance mappings. The conventional MDE algorithm[17] maps the desired value to the closest value produced by the SLM. Alternatively, we propose a modified MDE (mMDE) in which the desired value is first mapped to the closest value that can be pseudorandom encoded, and then the mapped value is encoded to a modulation value by PRE.

In this paper we will show, by using both computer simulations and experiments with a phase-only SLM, that the proposed modified blended encoding algorithm generally outperforms the earlier blended algorithm in terms of two metrics that describe fidelity of the reconstruction (the ratio of intensity of the desired portions of the diffraction pattern to peak background noise and the relative error in intensity between the desired and actual diffraction patterns). This demonstration is the primary objective of this paper. One secondary objective is to suggest how blended encoding algorithms can be developed for a variety of SLM modulation characteristics. Our approach is to develop blended algorithms for several different modulation characteristics. Another secondary objective is to provide a comprehensive comparison of the performance of various encoding algorithms developed to date. This is achieved by encoding an identical desired function for each algorithm and for each value of the scaling parameter γ.

Section 2 reviews the development of the proposed encoding algorithm and presents general background that is used to develop the new algorithms. Section 3 defines the modulation characteristics and the algorithms used in the study. Section 4 reports the results of the simulation study, and Section 5 presents the experimental results.

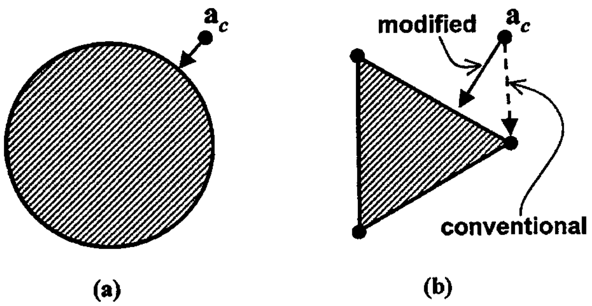


Fig. 1. Modulation characteristics for which the minimum-distance mapping to the modulation characteristic and to the encoding range (striped regions) of the PRE algorithm are (a) identical and (b) different. In (a) the modulation characteristic is a circle, and in (b) the modulation characteristic is the three dots, one at each apex of the triangle.

2. BACKGROUND ON AND REVIEW OF ENCODING

A. Review of Pixel-Oriented Encoding Algorithms

The design methods considered in this paper can be classified as pixel-oriented encoding, since each pixel represents a desired complex value independent of the values represented by all other pixels in the SLM.[18] Pixel-oriented encoding is a special case of point-oriented encoding. In traditional point-oriented encoding methods, the desired complex-valued function is modulated onto a carrier of spatial frequency that exceeds the space–bandwidth product (SBWP) of the desired complex function.[11,12,19] Therefore these methods require SLM's with SBWP's that exceed the SBWP of the desired signal. However, in pixel-oriented encoding the SBWP of the signal can be as large as that of the SLM as a result of the one-for-one mapping between each desired complex value and the modulation value of each corresponding pixel. Thus pixel-oriented encoding has an advantage over traditional point-oriented encoding, and also group-oriented encoding methods,[9,11,12] when the SLM has a small number of pixels, as is the case for most of the electrically addressable SLM's that are available today.

There appear to be two approaches to pixel-oriented encoding. One approach is to map each desired complex value to the closest available modulation value produced by its corresponding pixel.[17] For continuous-value phase-only SLM's this prescription leads to a unique mapping in which the amplitude of each value is set to unity and the mapped phase is identical with the desired phase. That is, MDE for the continuous-value phase-only SLM reduces to the well-known kinoform[20] or phase-only filter.[21]

The second encoding approach, PRE,[16,22] rather than selecting the closest available modulation value, selects one modulation value from a range of possible values by using a computer-generated random (i.e., pseudorandom) number. The statistical properties of the random-number generator are designed so that the average modulation value is identical with the desired complex value. The diffraction pattern produced by this transmittance function has an average intensity that is identical with the desired diffraction pattern plus a noise background. The diffraction efficiency η of the pseudorandom-encoded function is identical with that of the desired fully complex function. The remaining energy $1 - \eta$ is either scattered into the noise background for phase-only SLM's or scattered into noise and absorbed if the SLM is non-phase-only.

PRE differs from MDE in that MDE always maps the desired value to the closest available value on the modulation characteristic while PRE maps the desired value to closer modulation values with greater relative frequency than to modulation values that are farther away. For quantized modulation characteristics the encoding algorithms are analogous to the numerical rounding of floating-point numbers. MDE is analogous to nearest-integer rounding, while PRE corresponds to rounding to the nearest integer most frequently and to the furthest integer least frequently according to a random selection process. MDE and PRE are illustrated in Fig. 2 for tri-phase modulation (with modulation values $\mathbf{a}_{m1}$, $\mathbf{a}_{m2}$, and

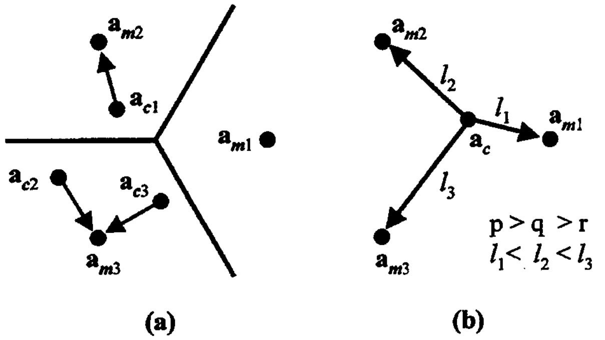


Fig. 2. Comparison of (a) the MDE algorithm with (b) the PRE algorithm for a tri-phase phase-only modulation characteristic.

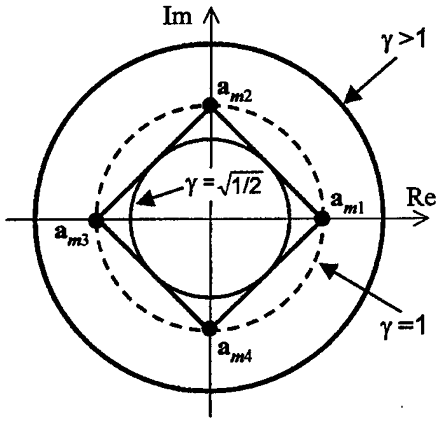


Fig. 3. Illustration of the encoding range and the fully complex encoding range, and their relationship to the scaling parameter γ for a quad-phase modulation characteristic.

$\mathbf{a}_{m3}$). For MDE [Fig. 2(a)], mapping to the closest value of modulation divides the complex plane into three decision regions separated by the three lines. For PRE [Fig. 2(b)] the modulation values $\mathbf{a}_{m1}$, $\mathbf{a}_{m2}$, and $\mathbf{a}_{m3}$ are randomly selected with the relative frequencies/probabilities p, q, and r (which are inversely proportional to l_1, l_2, and l_3, the distances between the desired value $\mathbf{a}_c$ and the modulator values). The complete mathematical specification of these PRE and MDE algorithms for tri-phase modulation and various other modulation characteristics are given in Section 3. The PRE algorithms for quantized modulation were originally derived and compared with MDE algorithms for quantized modulation in Ref. 22.

B. Review of Reduced-Parameter Suboptimal Design Methods

The earliest applications of design optimization using a few parameters appear in the work of Farn and Goodman[23] and in Juday[17,24] on the design of single-object correlation filters for limited-modulation-range SLM's.[17,23,24] Reference 17 presents this problem in its most general form. A fully complex filter is specified that optimizes a given performance metric. The filter has the complex-valued free parameter $\Gamma = \gamma \exp(j\beta)$, which scales the amplitude of the desired function by γ and rotates the phase by β. The desired function is encoded by the MDE algorithm for different values of γ and β. The optimal values γ^* and β^* minimize the sum of squares error between the desired function and the encoded values. There are no other free parameters for the single-object matched filter, and thus the design is optimal. However, more recent studies have reported suboptimal searches over these two parameters for the design of composite pattern recognition filters[25] and spot array generators.[26] These studies predated and stimulated the development of the first algorithms that blend MDE and PRE to various degrees as a function of the free parameter γ.[14,15]

In specific cases searches over one or both of the free parameters can be avoided. For specific modulation characteristics the encoding algorithm can be independent of γ and/or β. For instance, in Fig. 2(a), the sum of squares error for MDE is independent of γ but dependent on β. As stated in Section 1, for continuous-value phase-only SLM's the MDE algorithm reduces to the classical kinoform, and thus no search is required at all. Also, the distribution of the desired values over the complex plane can make the optimization insensitive to the variation in γ or β. For instance, in Fig. 2, if the desired complex values are uniformly distributed in magnitude and phase, then there is essentially no dependence on the value of β. This observation is used in the present study to perform single-parameter searches over γ for both MDE and blended encoding.

C. Pseudorandom Encoding Range and Fully Complex Encoding Range

Another reason for the development of blended algorithms is that while MDE algorithms can encode complex values of any magnitude, the PRE algorithm cannot.[27] This is because in PRE the desired value is encoded so that its complex value is equivalent to an average of the available modulation values, and the average is thus constrained to lie between the modulation values. A procedure for evaluating the range over the complex plane that can be encoded by a given PRE algorithm is developed in Ref. 27. Ranges for the modulation characteristics considered in this paper are shown in Figs. 1 and 3. For continuous phase-only modulation [Fig. 1(a)], the PRE range is the interior of the unit circle. Figure 1(b) shows the range for three-value quantized phase modulation. The encoding for the PRE algorithm is the triangular region that is enclosed by the line segments connecting the three values of modulation. Similarly, for a four-value quantized phase modulation (Fig. 3), the encoding range is the square and its interior, which is defined by the line segments connecting the modulation values.

Note that in Fig. 3 the desired complex function can be normalized so that its values are contained within a circle of radius $\gamma = \sqrt{1/2}$. We refer to this as the fully complex encoding range for the quantized PRE algorithm. (For individual functions for which the distribution of complex values is noncircular, the fully complex range can approach $\gamma = 1$. However, we apply this definition not to individual functions but rather to the set of all functions of interest.) Also note that for phase-only modulation the encoding range and the fully complex range are identical (the region enclosed by $\gamma = 1$ in Fig. 3) and that they enclose a larger area of the complex plane than does the quantized PRE algorithm. In this paper the encoding range is increased by blending PRE algorithms with algorithms that do not have limited encoding ranges. As shown in Sections 4 and 5, the diffraction efficiency is increased and the fidelity is optimized for fully complex encoding ranges (as designated by the scaling parameter γ) that exceed the encoding range of PRE alone.

3. DESIGN OF THE STUDY

A. Modulation Characteristics

The modulation characteristics considered in this study are illustrated in Fig. 4. Three of the characteristics [(a)–(c)] are phase only: (a) continuous, (b) three phases uniformly spaced around the unit circle, and (c) four uniformly spaced phases. Adding an additional zero value to each characteristic gives the bi-amplitude modulation characteristics [(d)–(f)]. We will refer to these modulation characteristics by using the descriptive terms tri-phase and quad-phase. Also, we use the terms bi-amplitude phase and phase-only to distinguish between modulation characteristics that have or do not have a zero value.

B. Encoding Algorithms

The implementation and the theory of PRE and MDE have been presented in the publications reviewed in Sections 1 and 2. We present only the details necessary to permit others to understand and to reproduce the results presented in Sections 4 and 5. As an aid to the reader, each of the specific algorithms studied is presented in the figures. We begin with the less-involved algorithms for the continuous modulation characteristics [Figs. 4(a) and 4(d)] and proceed through increasingly involved algorithms for tri-phase [Figs. 4(b) and 4(e)] and quad-phase [Figs. 4(c) and 4(f)] characteristics.

C. Encoding Algorithms for Continuous Spatial Light Modulators

Figure 5(a) illustrates MDE for a phase-only SLM. As mentioned in Subsection 2.B because of circular symmetry of the modulation characteristic the desired fully complex function [illustrated by the values $\mathbf{a}_{c1}$ and $\mathbf{a}_{c2}$ in Fig. 5(a)] can be scaled by an arbitrary complex number Γ and the encoding still maps to the unit circle identically. However, with the addition of a zero value of modulation MDE for the bi-amplitude phase modulator, the mapping becomes more involved in two respects: (1) While the mapping is still along radial lines, there is now a threshold level [dashed curve in Fig. 5(d)]. Values less than radius 1/2 are closer to the origin than to unity and therefore map to the origin. (2) Because of the threshold the mapping now depends on the magnitude of the scaling parameter γ. For $\gamma = 0$ all the desired complex values map to $\mathbf{a}_0$, and for $\gamma = \infty$ the complex values map to the unit circle, which is identical with the mapping in Fig. 5(a).

Our convention for reporting the value of the scaling parameter γ (for all encoding algorithms presented) is as follows: The desired complex function consists of N samples $\mathbf{a}_{ci}$ at positions indexed by i from 1 to N. The complex values are normalized so that the maximum value of $|\mathbf{a}_{ci}|$ from the N samples is identical with γ. The value of γ that produces the best performance according to a given metric or cost function is usually written as γ^*.

Figure 5(b) illustrates PRE for phase-only SLM's. This particular algorithm was introduced in Ref. 22. The desired value $\mathbf{a}_{c1}$ is mapped to one of two modulation values that are 180° apart. For each pixel transmittance

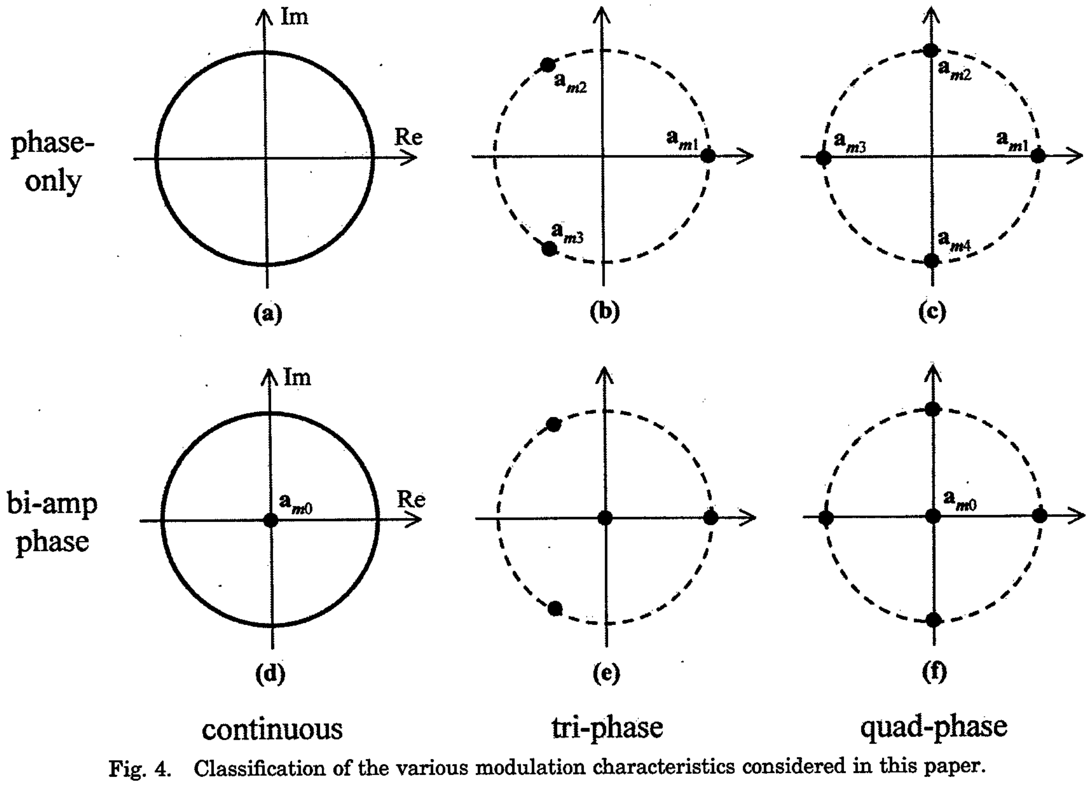


Fig. 4. Classification of the various modulation characteristics considered in this paper.

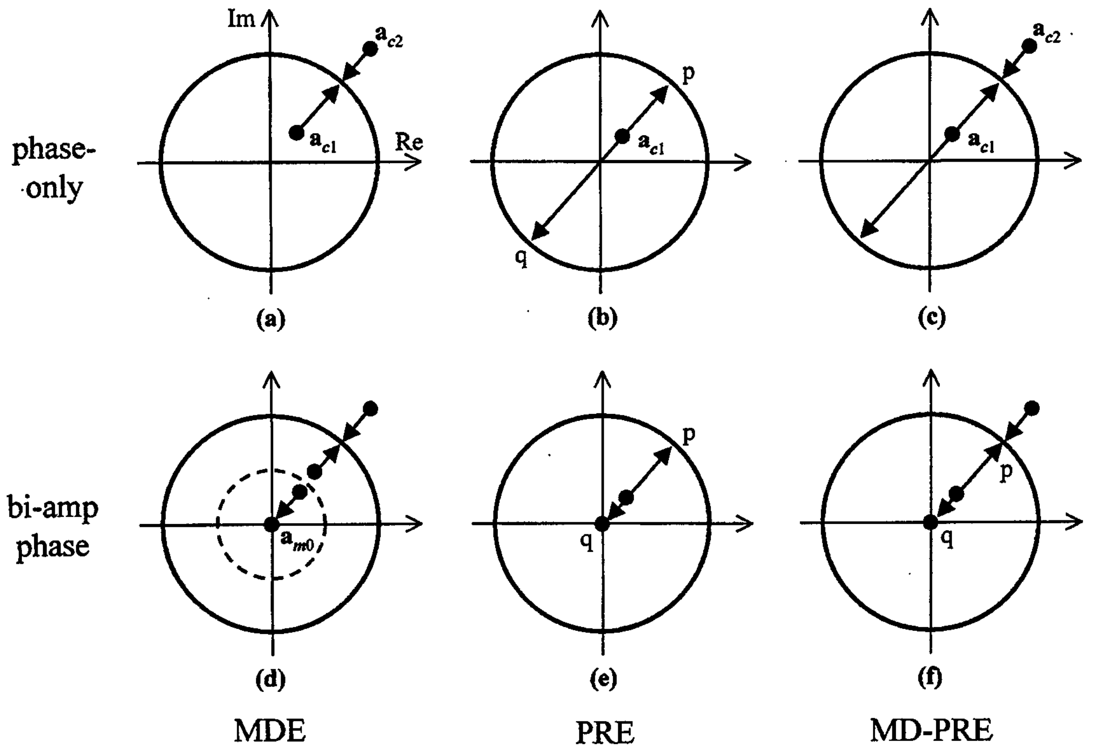


Fig. 5. Illustration of the individual MDE and PRE algorithms together with their blending for continuous modulation characteristics.

$\mathbf{a}_{ci}$, the probability of selecting the modulation value that is closer to the desired value is

$$p_i = (1 + |\mathbf{a}_{ci}|)/2, \tag{1}$$

and the probability of selecting the alternative value is $q_i = 1 - p_i$. With these values of probability, the encoding formula is

$$\mathbf{a}_i = \begin{cases} \exp[j \arg(\mathbf{a}_{ci})] & \text{if } 0 \leq s_i < p_i \\ -\exp[j \arg(\mathbf{a}_{ci})] & \text{if } p_i \leq s_i \leq 1 \end{cases}, \tag{2}$$

where $\mathbf{a}_i$ is the actual modulation selected for the ith modulator pixel and s_i is a computer-generated random number uniformly distributed between 0 and 1. To reduce encoding errors, one usually tries to make the value of γ as large as possible.[27] For phase-only SLM's this corresponds to $\gamma = 1$. For values of $\gamma > 1$ the complex values that exceed unity cannot be pseudorandom encoded. These values can be encoded by MDE, which leads to the blended minimum-distance and pseudorandom encoding algorithm (MD-PRE) illustrated in Fig. 5(c).

The PRE algorithm for bi-amplitude phase modulation is illustrated in Fig. 5(e). The probability is

$$p_i = |\mathbf{a}_{ci}|, \tag{3}$$

and $q_i = 1 - p_i$. The encoding formula is

$$\mathbf{a}_i = \begin{cases} \exp[j \arg(\mathbf{a}_{ci})] & \text{if } 0 \leq s_i < p_i \\ 0 & \text{if } p_i \leq s_i \leq 1 \end{cases}. \tag{4}$$

The MD-PRE algorithm for bi-amplitude phase modulation [Fig. 5(f)] uses the PRE algorithm for encoding values inside the unit circle and phase-only MDE for encoding values outside the unit circle.

D. Ternary Pseudorandom Encoding

The encoding formula for ternary-valued modulation[22] is presented here in general form because it is the basis for the PRE algorithms for all the quantized SLM's considered in this study [Figs. 4(b), 4(c), 4(e), and 4(f)]. The ternary PRE algorithm can be specified for any three modulation values $\mathbf{a}_{m1}$, $\mathbf{a}_{m2}$, and $\mathbf{a}_{m3}$ as long as they do not lie on a common line. The encoding formula is

$$\mathbf{a}_i = \begin{cases} \mathbf{a}_{m1} & \text{if } 0 \leq s_i < p_i \\ \mathbf{a}_{m2} & \text{if } p_i \leq s_i < 1 - r_i, \\ \mathbf{a}_{m3} & \text{if } 1 - r_i \leq s_i \leq 1 \end{cases} \tag{5}$$

where p_i is the probability of selecting $\mathbf{a}_{m1}$, q_i is the probability of selecting $\mathbf{a}_{m2}$, and r_i is the probability of selecting $\mathbf{a}_{m3}$. As in Subsection 3.C, s_i is a random number drawn from the uniform probability distribution. The three probabilities are found by solving

$$\begin{pmatrix} \mathrm{Re}(\mathbf{a}_{ci}) \\ \mathrm{Im}(\mathbf{a}_{ci}) \\ 1 \end{pmatrix} = \begin{bmatrix} \mathrm{Re}(\mathbf{a}_{m1}) & \mathrm{Re}(\mathbf{a}_{m2}) & \mathrm{Re}(\mathbf{a}_{m3}) \\ \mathrm{Im}(\mathbf{a}_{m1}) & \mathrm{Im}(\mathbf{a}_{m2}) & \mathrm{Im}(\mathbf{a}_{m3}) \\ 1 & 1 & 1 \end{bmatrix} \begin{pmatrix} p_i \\ q_i \\ r_i \end{pmatrix}, \tag{6}$$

where $\mathbf{a}_{ci}$ is the desired complex value that is encoded. For quantized SLM's that have more than three modulation values, the PRE algorithm is developed by using Eqs. (5) and (6) with various groups of three modulation values to encode various regions of the complex plane.

E. Encoding Algorithms for Quantized Phase-Only Spatial Light Modulators

Figure 6 illustrates how the individual MDE and PRE algorithms are combined into the MD-PRE and modified

MD-PRE (mMD-PRE) algorithms for both tri-phase and quad-phase modulation characteristics.

For MDE on tri-phase SLM's, the complex plane is divided into three decision regions [Fig. 6(a)], and desired values in a particular region are mapped to the modulation value in that region. For PRE on tri-phase SLM's [Fig. 6(b)], the modulation values used in Eqs. (5) and (6) are $\mathbf{a}_{m1} = 1$, $\mathbf{a}_{m2} = \exp(j2\pi/3)$, and $\mathbf{a}_{m3} = \exp(-j2\pi/3)$. Values in the interior of the triangle in Fig. 6(b) can be pseudorandom encoded, and the inscribed circle (dashed curve), which is of radius $\gamma = 0.5$, represents the fully complex range for this PRE algorithm. The MD-PRE blénded algorithm uses PRE for desired values on and inside the triangle of Fig. 6(b), and it uses the MDE decision regions of Fig. 6(a) for values outside the triangle. The PRE and MDE regions for the blended algorithms are labeled in Fig. 6(c).

As in Fig. 1(b), MD-PRE can be modified to the mMD-PRE algorithm by mapping desired values that are outside the PRE range to the closest values on the boundary of the PRE range. Then the mapped value is encoded by the PRE algorithm. We will refer to the mapping of values by this prescription as modified MDE (mMDE). The mMDE regions are identified in Fig. 6(d). The regions identified as MDE in Fig. 6(d) are also encoded by the mMDE prescription; however, mMDE for these regions is identical with MDE.

The mMD-PRE for the quad-phase SLM is developed in a similar manner to that described for tri-phase encoding. Figures 6(e)–6(h) illustrate the corresponding succession of steps for the quad-phase SLM. Note that for the quad-phase PRE algorithm the fully complex range becomes $\gamma = \sqrt{1/2}$, as indicated by the dashed curve in Fig. 6(f). Also note that Fig. 6(f) distinguishes between two regions in the encoding range of the PRE algorithm. For each region a tri-phase PRE algorithm is used. If $\mathbf{a}_{ci}$ is in region I, then it is encoded by using the modulation values $\mathbf{a}_{m1} = 1$, $\mathbf{a}_{m2} = j$, and $\mathbf{a}_{m3} = -1$, and if $\mathbf{a}_{ci}$ is in region II, then it is encoded by using $\mathbf{a}_{m1} = 1$, $\mathbf{a}_{m4} = -j$, and $\mathbf{a}_{m3} = -1$. The encoding formula can be written as

$$\mathbf{a}_i = \begin{cases} 1 & \text{if } 0 \leqslant s_i < p_i \\ \pm j & \text{if } p_i \leqslant s_i < 1 - r_i, \\ -1 & \text{if } 1 - r_i \leqslant s_i \leqslant 1 \end{cases} \qquad (7)$$

where in the second line j is used if $\mathbf{a}_{ci}$ is in region I and $-j$ is used if $\mathbf{a}_{ci}$ is in region II. The values of probability used in Eq. (7) are determined by solving the equation

$$\begin{pmatrix} \mathrm{Re}(\mathbf{a}_{ci}) \\ \mathrm{Im}(\mathbf{a}_{ci}) \\ 1 \end{pmatrix} = \begin{bmatrix} 1 & 0 & -1 \\ 0 & \pm 1 & 0 \\ 1 & 1 & 1 \end{bmatrix} \begin{pmatrix} p_i \\ q_i \\ r_i \end{pmatrix}, \qquad (8)$$

where $+1$ is used if $\mathbf{a}_{ci}$ is in region I and -1 is used if $\mathbf{a}_{ci}$ is in region II.

F. Encoding Algorithms for Quantized Biamplitude Phase Spatial Light Modulators

Figure 7 identifies the various encoding regions for PRE and MDE with the addition of the modulation value $\mathbf{a}_{m0} = 0$. For both tri-phase MDE [Fig. 7(a)] and quad-phase MDE [Fig. 7(c)], one additional decision region is formed. For PRE on the tri-phase SLM, there are three regions, each of which is pseudorandom encoded by using the

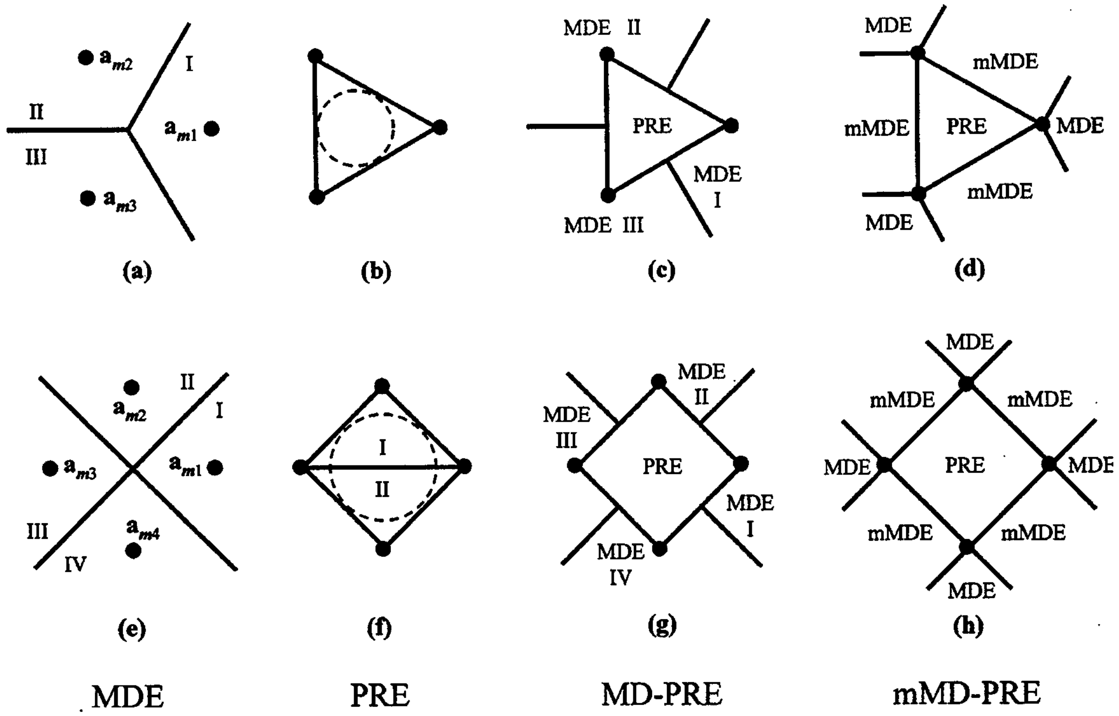


Fig. 6. Illustration of the individual MDE and PRE algorithms together with their minimum-distance and modified-minimum-distance blendings for [(a)–(d)] tri-phase and [(e)–(h)] quad-phase phase-only modulation characteristics. Parts (a) and (e) identify the decision regions for MDE. Parts (b) and (f) show the encoding ranges for the PRE algorithms together with the fully complex ranges, which are bounded by each dashed circle. Part (f) also indicates that there are two regions. Each triangular region is encoded by Eqs. (7) and (8) with use of the three modulation values at the corners of the corresponding regions.

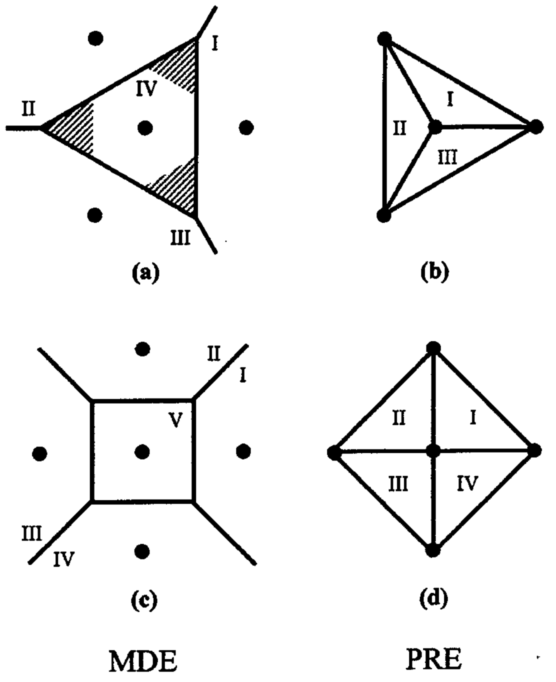


Fig. 7. Illustration of the individual MDE and PRE algorithms for bi-amplitude modulation characteristics: (a) MDE and (b) PRE for tri-phase SLM's; (c) MDE and (d) PRE for quad-phase SLM's. Parts (a) and (c) show the decision regions for the MDE algorithms. Parts (b) and (d) show the individual subregions that are each encoded by using ternary PRE. The striped areas of region IV in (a) are outside the encoding range for PRE in (b). Therefore the MD-PRE blending of (a) and (b) requires that values in the striped areas be mapped to zero according to the MDE algorithm.

modulation values $\mathbf{a}_{m0} = 0$ and two of the three other values $\mathbf{a}_{m1}$, $\mathbf{a}_{m2}$, and $\mathbf{a}_{m3}$ in Fig. 7(b). Similarly, for PRE with a quad-phase SLM there are four regions, each of which is pseudorandom encoded by using Eqs. (5) and (6) with $\mathbf{a}_{m0}$ and two of the four other values $\mathbf{a}_{m1}$, $\mathbf{a}_{m2}$, $\mathbf{a}_{m3}$, and $\mathbf{a}_{m4}$ in Fig. 7(d).

Even though the PRE algorithms used for the quantized bi-amplitude SLM's are different from those used for the quantized phase-only SLM's, the encoding ranges of the PRE algorithms are identical. This leads to the PRE and MDE regions for the blended algorithms being identical with those for the phase-only SLM's [Figs. 6(c), 6(d), 6(g), and 6(h)], with one exception [Fig. 6(c)]. From Figs. 7(c) and 7(d), it can be seen that MDE region V is entirely contained inside the PRE encoding range, and thus this region is always encoded by PRE and is never encoded by MDE. Therefore Figs. 6(g) and 6(h) apply to the bi-amplitude SLM as well. However, Figs. 7(a) and 7(b) show that some of MDE region IV (the three striped regions) are outside the encoding range for PRE. Therefore the three striped regions should be added onto Fig. 6(c) to properly describe the encoding regions for MD-PRE for the tri-phase bi-amplitude SLM. For mMD-PRE the mMDE regions take precedence over the MDE IV regions, and thus Fig. 6(d) describes the encoding range for both the phase-only and the bi-amplitude SLM.

In passing, we note that for MD-PRE on tri-phase bi-amplitude SLM's there is a dramatic difference between encoding a value in the PRE region and in the MDE IV regions that are outside the PRE region. Desired values in the PRE region are mapped on a percentage basis to one of the three closest SLM values, while values in the MDE IV region are always mapped to zero. This algorithm has led to the somewhat paradoxical result that some values further from zero are mapped to zero more frequently than other values that are closer to zero.

G. Specification of the Desired Function To Be Encoded

The desired function that is encoded is written in the form

$$\mathbf{a}_c(x, y) = \sum_{k=1}^{7} \exp(j\theta_k)\exp(j2\pi kx) \times \sum_{l=1}^{7} \exp(j\theta_l)\exp(j2\pi ly), \qquad (9)$$

where θ_k are the phases specified by Krackhardt *et al.* for a maximum-diffraction-efficiency, phase-only 1 × 7 spot array.[28] Equation (9), which is periodic, is sampled to produce a 32 × 32 unit cell of complex values, and from this a 4 × 4 array of cells is adjoined to produce the array of 128 × 128 desired complex values.

Neither PRE nor blended encoding requires that the desired function be designed by optimization. Nor were optimized functions encoded in Refs. 14 and 15. However, it is useful to use Eq. (9) because this function and its performance are well-known and because it provides information that relates the performance of encoding procedures to the performance of optimized designs.

Since Eq. (9) is periodic, one might also wonder whether periodic functions have a performance advantage over nonperiodic functions. Reference 15 is the only study of blended algorithms that uses a nonperiodic function. However, several of our studies on PRE alone have identified that the SBWP of the desired function, the diffraction efficiency of the desired function [see Eq. (26) in Ref. 22] and the mean squared distance between the desired function values and the modulation values critically control performance.[22] References 8, 16, and 18, which include simulated and experimental demonstrations using nonperiodic functions, and Ref. 22, which uses periodic functions, all demonstrate similar dependence on these parameters that define the properties of the function.

The encoding of the optimized function suggests that, in addition to design, the encoding algorithms could also be used to remap an optimized design from one type of modulation characteristic into another. A specific application of remapping would be to use the encoding algorithms for quantized modulation to quantize a continuous-value phase-only diffractive optical element design.

H. Simulation Procedures and Definition of the Performance Metrics

The far-field diffracted intensity patterns are simulated by fast-Fourier-transforming the encoded values $\mathbf{a}_i$ and then squaring the magnitude for each of the pseudorandom and nonrandom encodings. For all metrics except diffraction efficiency, the 128 × 128 array is placed in a

512×512 array of zeros that is fast Fourier transformed. For diffraction efficiency the 128×128 array is fast Fourier transformed directly. The diffraction efficiency η is simply the sum of the intensities of the 49 spots divided by the sum of all intensities in the 128×128 diffraction pattern. For bi-amplitude modulation characteristics the energy absorption in the modulator plane also needs to be accounted for.[14] Therefore the ratio of desired energy to total energy in the diffraction pattern is multiplied by the ratio of unity-transmittance pixels to the total number of SLM pixels. Nonuniformity of the peaks (NU) is calculated as the standard deviation of the peak intensities of the 49 spots divided by the average spot intensity. Signal-to-peak-noise ratio (SPR) is the ratio of the average peak intensity of the spots to the maximum noise peak of the 512×512 pattern, excluding the square region that contains the 7×7 spot array. Signal-to-noise ratio (SNR) is the ratio of the average intensity of the peak values of each of the 49 spots divided by the average intensity outside the square region containing the 7×7 spot array. SNR is reported for completeness and to provide continuity with the results and the theory on the performance of ternary PRE that was presented in Ref. 22. However, in Section 4 we provide little discussion of the SNR results because SNR does not well characterize the noise in MDE and MD-PRE, which, rather than being white, is localized to a small number of large noise spikes. The calculation of the various metrics from experimental measurements is described in Section 5.

In addition to describing the specific encoding algorithms that are to be evaluated in this study, we hope that our development of these algorithms may serve as examples and suggest how blended encoding algorithms can be developed for the myriad of possible modulation characteristics.

4. SIMULATION RESULTS

This section compares the performance of PRE, MDE, MDE-PRE, and mMDE-PRE algorithms for various modulation characteristics in terms of SPR, NU, and η as a function of the blending/scaling parameter γ and also at selected optimal values of γ^*.

For each encoding performed in this study, the identical desired function $\mathbf{a}_{ci}$ and the 128×128 array of uniform random numbers s_i are used. Using the same random numbers is important because each new set of random numbers used in encoding can affect the value of the performance metrics. However, even using an identical array of random numbers still causes fluctuations in the performance curves. These fluctuations can be reduced by performing the same encoding algorithm multiple times with several sets of random numbers and then averaging together the performance metrics.[8] However, the trends in the performance curves are adequately evident for the purpose of comparing the performance advantages of one algorithm with those of another.

The detailed performance results for the various SLM types are reported below. The first set of results is for continuous SLM's. While there is no distinction between MD-PRE and mMDE-PRE for these characteristics, the results for continuous SLM's provide the clearest demonstration of the improvements that are due to blending and they also provide a baseline against which to compare the performance when the phase characteristic is coarsely quantized.

A. Results for Continuous Spatial Light Modulators

Figure 8 shows the performance as a function of γ for the encoding of the identical function on phase-only and bi-amplitude phase SLM's. MD-PRE for both SLM types is presented together with MDE for the bi-amplitude SLM (which for $\gamma = \infty$ is equivalent to MDE for the phase-only SLM). For each SPR and NU curve, there is a particular value of $1 < \gamma \ll \infty$ for which the performance metric is optimal. The performance metrics for each algorithm when SPR is maximum are reported in Table 1. Since NU is fairly flat in the vicinity of peak SPR, these additional data are not presented. Comparing the curves and

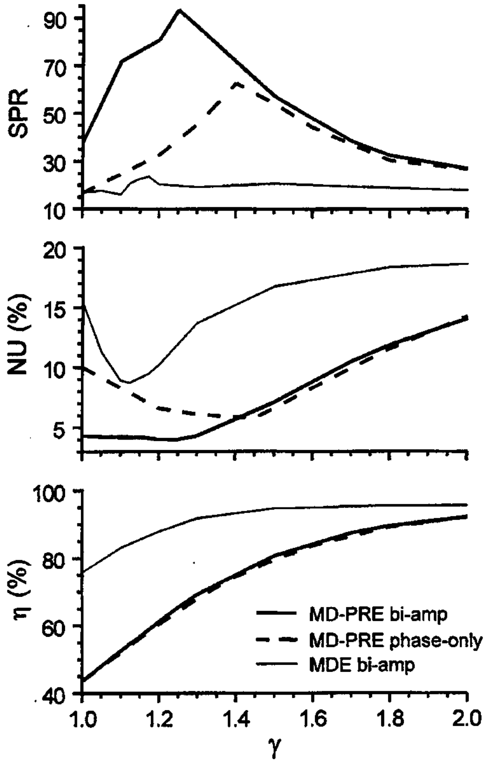


Fig. 8. Simulated performance of blended algorithms as a function of the blending parameter for phase-only and bi-amplitude phase modulation characteristics.

Table 1. Best Performance of Encoding Algorithms for Continuous SLM's

Continuous		Simulation (Experiment)			
Algorithm	γ^*	η (%)	SNR	SPR	NU (%)
Phase-only					
PRE	1.00	44(44)	254 (224)	17(13)	10(15)
MD-PRE	1.40	75(70)	977 (743)	63(43)	6(15)
MDE	∞	96(94)	2263(2135)	20(15)	19(23)
Bi-amplitude					
PRE	1.00	43	685	37	4
MD-PRE	1.25	65	1524	93	4
MDE	1.20	88	7093	20	10

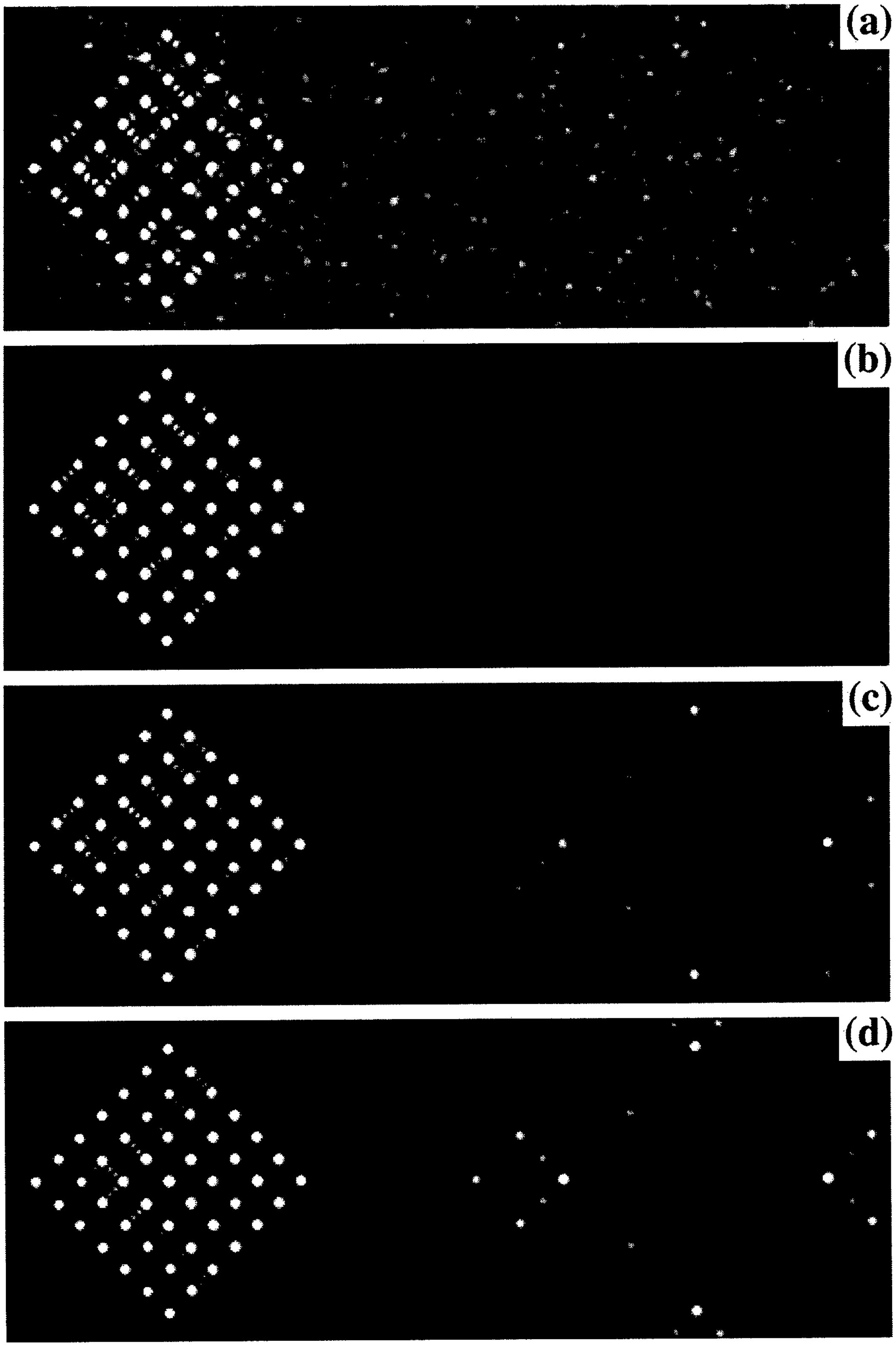


Fig. 9. Simulated far-field intensity patterns of the tri-phase phase-only SLM for (a) PRE, (b) mMD-PRE, (c) MD-PRE, and (d) MDE. The images show intensity with a linear gray scale. To bring out the background noise, the maximum gray-scale value (full white) is 30% of the average intensity of the 49 spots.

table entries with each other, we can also see that MD-PRE for bi-amplitude SLM's outperforms MD-PRE for phase-only SLM's. Both algorithms outperform MDE in SPR and NU. Clearly, MDE produces greater diffraction efficiency; however, the diffraction efficiency for the MD-PRE algorithms can exceed 80% (near $\gamma = 1.6$) and still outperform the best MDE in terms of SPR and NU. The trends in these performance curves are similar to that observed in Ref. 14, where a nonoptimized, lower-diffraction-efficiency function was encoded.

B. Results for Quantized Phase-Only Spatial Light Modulators

The characteristics of the various algorithms and their influence on performance can be appreciated by considering the simulated diffraction patterns of Fig. 9. The values of γ^* used for each type of encoding are reported in Table 2 together with the tabulated performance metrics. The gray scale in Fig. 9 has been set to bring out the structure of the background noise. The background for PRE is a relatively bright speckle pattern [Fig. 9(a)], while the background for MDE is a much different pattern of noise spikes at harmonically related spatial frequencies [Fig. 9(d)]. The background for MD-PRE [Fig. 9(c)] also contains noise spikes having a similar spatial distribution of noise to that for MDE but that are not as bright as those for MDE. There is also a speckle pattern that is quite faint. The background for mMD-PRE [Fig. 9(b)] contains a speckle background that is slightly brighter than the speckle pattern for MD-PRE but that is much less bright than the patterned noise spikes for MD-PRE. There are even patterned noise spikes in Fig. 9(b), but they are faint and obscured to a large degree by the speckle pattern. Figure 9 has been used to show how blending trades off between the background noise properties of PRE and MDE. The cross sections in Fig. 10 of the intensity patterns allow a more quantitative comparison of the performance of the four encoding algorithms. Figure 10 makes clear that it is the appearance of a few very large noise spikes that leads to the low values of SPR for MDE and MD-PRE. Figure 10 also provides a visual comparison of uniformity of the spot arrays. While the mMD-PRE is the most uniform of the four cross sections, the differences are best appreciated by considering the values of NU reported in Table 2 for the uniformity of all 49 spots. The same visual and qualitative distinctions for the four diffraction patterns can be made for encoding with the more finely quantized modulation characteristics [Figs. 4(c), 4(e), and 4(f)] considered in this study. Since noise spikes and the nonuniformity are generally lower, these differences are harder to see and they provide no additional insight into the properties of the encoding algorithms. For this reason the algorithms are compared in terms of their performance metrics in the remainder of the paper.

The performance of the blended encoding algorithms as a function of γ is given in Fig. 11. While the curves are noisier than the continuous curves in Fig. 8, it can be seen that for each SPR curve the maximum value is found for a specific value of γ^* corresponding to a specific blending of PRE and MDE (or mMDE). We have never found a case in which either pure PRE or pure MDE produced a better performance than the blended results. Similarly, for each NU curve, the minimum value corresponds to a specific value of the blending parameter γ^*. Of most significance to this study is that the mMD-PRE curves always attain larger values of SPR and lower values of NU than the corresponding MD-PRE curves. This is true despite the fact that MD-PRE has the larger diffraction efficiency. Rather than reporting the best SPR and the best NU, Table 2 reports the performance for the best overall combination of SPR and NU (as based on empirical judgment rather than cost function). The selection of the best value of γ^* is not especially critical because NU (or SPR) is slowly varying near its local minimum (or maximum).

Table 2. Best Performance of Encoding Algorithms for Phase-Only SLM's

Phase-Only		Simulation (Experiment)			
Algorithm	γ^*	η (%)	SNR	SPR	NU (%)
Tri-phase					
PRE	0.50	11(11)	40(37)	3(3)	20(23)
MD-PRE	0.80	32(32)	147(152)	2(6)	15(18)
mMD-PRE	0.95	31(28)	147(130)	11(9)	12(15)
MDE	∞	66(57)	411(429)	1(1)	22(25)
Quad-phase					
PRE	0.71	22(18)	93(77)	8(6)	11(14)
MD-PRE	0.95	40(39)	220(196)	13(7)	11(15)
mMD-PRE	1.11	47(45)	290(274)	22(18)	8(14)
MDE	∞	78(77)	1071(949)	4(7)	21(25)

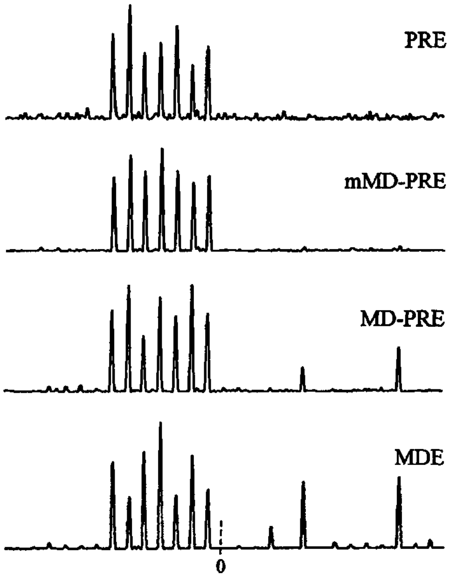


Fig. 10. Cross sections of the far-field intensity patterns of the tri-phase phase-only SLM from Fig. 9. The cross section is taken across the diagonal of the 7 × 7 spot array and through the optical axis (indicated by the dashed vertical line). The traces are normalized so that the average intensity of each spot array is of identical vertical length on each plot.

C. Results for Quantized Biamplitude Phase Spatial Light Modulators

Figure 12 and Table 3 report these results. Similar trends to those noted for the quantized phase-only SLM's are observed for the biamplitude SLM's. Once again each curve demonstrates that there is a particular degree of blending that produces the best fidelity as measured by SPR or NU. Also, the largest value of SPR for mMD-PRE is always greater than the largest value of SPR for MD-PRE. Similarly, the smallest value of NU for mMD-PRE is always smaller than the smallest value of NU for MD-PRE. The diffraction efficiency for the tri-phase encodings shows that the MD-PRE actually has lower diffraction efficiency than mMD-PRE for $\gamma \lesssim 1$. This reflects the fact that many of the values in the MDE region IV [specifically, the striped regions of Fig. 7(a)] are being mapped to zero.

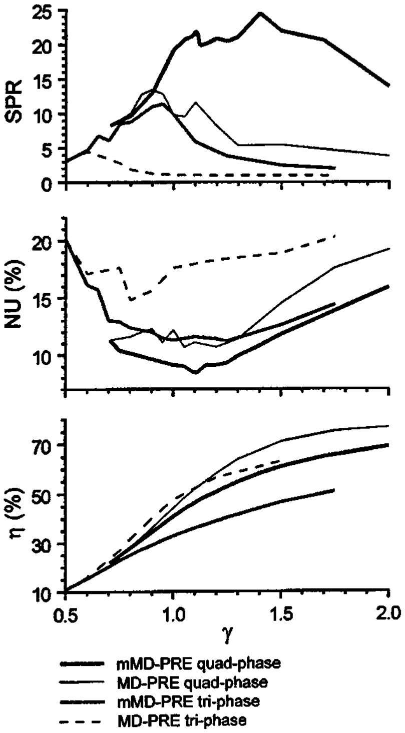


Fig. 11. Simulated performance of blended algorithms as a function of the blending parameter for quantized-phase phase-only modulation characteristics.

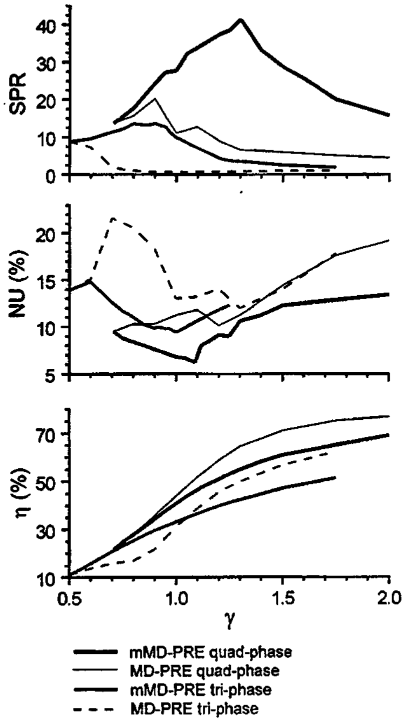


Fig. 12. Simulated performance of blended algorithms as a function of the blending parameter for quantized-phase bi-amplitude phase modulation characteristics.

Table 3. Best Performance of Encoding Algorithms for Bi-Amplitude SLM's

Bi-Amplitude		Simulation			
Algorithm	γ^*	η (%)	SNR	SPR	NU (%)
Tri-phase					
PRE	0.50	11	88	9	14
mMD-PRE	0.90	30	170	14	10
MD-PRE	1.30	50	309	1	12
MDE	1.50	58	366	1	13
Quad-phase					
PRE	0.71	22	197	14	10
mMD-PRE	1.20	51	443	37	8
MD-PRE	1.20	59	606	9	10
MDE	1.20	65	938	5	11

Comparing the quantized biamplitude results with the quantized phase-only results shows that the extra zero-valued state markedly improves the fidelity measures. The diffraction efficiency curves in Figs. 11 and 12 are essentially identical as a function of γ except for the tri-phase MD-PRE curve [which differs because region IV in Fig. 7(a) extends outside the PRE region]. Also, the diffraction efficiency performance reported in Tables 2 and 3 depends directly on the value of γ^* required to optimize the fidelity metrics. Therefore the diffraction efficiency of blended algorithms on phase-only SLM's turns out sometimes to be higher and sometimes to be lower than the efficiency of blended algorithms on biamplitude SLM's.

5. EXPERIMENTAL RESULTS

A. Spatial Light Modulator Characterization and Experimental Procedures

A Boulder Nonlinear Systems Inc. (BNS) 128 × 128-pixel reflective SLM is used in our experiments. BNS normally sells this SLM with ferroelectric liquid crystal (LC). On request they will fill the cell with parallel aligned nematic LC, as was done for us and other groups as well.[29] The relation between voltage and phase modulation has been determined by two methods. One is an interferometric method based on Young's fringes.[30] The second uses the diffraction pattern of a random bi-phase distribution.[8]

For a perfect device the two methods should lead to the same phase characteristic; however, variations in phase response are known to occur across the device.[29] We find experimentally that using the results from the random bi-phase method for phases up to π and from the interferometric method for the range π–2π gives the best correspondence between the actual and an ideal phase-only SLM. For a frequency-doubled Nd-YVO$_4$ laser (532 nm), we found that a 2π range is produced with 80 electrically addressable gray-scale levels. However, because of the nonlinear transfer curve that is typical for LC SLM's, the phase levels are not equally spaced.

For the measurement of the spot arrays, the linearly polarized laser beam is spatially filtered and collimated. The reflective SLM is uniformly illuminated, with linearly polarized light oriented with the extraordinary axis of the LC. The light reflected from the SLM is collected by a Fourier transform lens, and the resulting Fraunhofer diffraction pattern (specifically the zero diffraction order of the SLM grating) is recorded with a Cohu 4915 CCD camera and attached National Instruments black-and-white frame grabber.

After any fixed background noise is subtracted off, the performance metrics are calculated as described in Subsection 2.H with the following exceptions: A noticeable spot, which is due to reflections from the cover glass of the SLM, is always present on the optical axis. It is omitted from all the calculations. The average background noise level is used in calculating not only SNR but also diffraction efficiency η. The average noise level is determined by adding the intensity in several regions (which excludes the undesired on-axis spot and which covers approximately 40% of the total area in the zero order) and then dividing by the area of these regions. The average of the maximum intensity of each of the desired spots is also calculated, and the ratio of the two averages gives SNR. The peak noise spike is found in the identical region that is used for the average noise level calculation. This value of peak noise is used in the calculation of SPR. For diffraction efficiency calculations the average noise level is multiplied by the total area of the zero-order region to estimate the noise energy. This, together with the energy in the desired spots, is considered to be the total energy for purposes of comparing the diffraction efficiency of the experiment with that of the theory.

While calculating efficiency in the above way does provide good agreement between experiment and theory, it is not representative of the physically true efficiency of the BNS SLM. What we find by using a power meter to measure incident and reflected light from the SLM (with all pixels set to gray-scale level 0) is that 1.7% of the energy appears in the on-axis spot, ~5.7% of the energy appears in all diffraction orders (as measured by reimaging the SLM onto the detector of the power meter), and a surprisingly low 0.9% of the energy appears in the zero-order diffraction pattern.

These measurements are aided by a slight lack of parallelism between some of the surfaces in the SLM, which causes the unmodulated spot from the cover glass of the SLM to become spatially resolved from the modulated spot at large distances from the SLM. When a linear phase ramp is programmed on the SLM, we observe that the modulated spot is translated with 95% of its energy present in the translated position and essentially no energy present in the original position. In the higher orders there is also a translated spot, but the unmodulated spot is undetectable. This leads to the conclusion that the unmodulated spot is from a continuous surface that has no spatially varying modulation. That is to say, there is no additional contribution to the unmodulated spot from the dead zones between the pixels. This is further supported by images of the SLM in the phase-only configuration that show dark lines between the pixels. However, in the amplitude-modulating configuration, when viewed through crossed polarizers, the dead zones are bright, which shows that they rotate the polarization.

Also, the losses cannot be attributed entirely to pixel fill factor. BNS quotes the pixel fill factor of a typical device as 60%, and we measured a 54% fill factor when viewing images of the SLM under incoherent illumination. Fill factors in this range indicate that between 29% and 36% of the modulated reflection should appear in the zero order. Therefore we conclude that there are losses of more than 1 order of magnitude in the LC cell.

B. Performance of the Encodings

The phase-only designs summarized in Tables 1 and 2 are implemented with the phase-only SLM, and the measured performance is reported in parentheses in the tables. The measured diffraction efficiencies and SNR are usually quite close, though somewhat less than the simulated values. There are larger deviations between the simulated and measured SPR and NU, with measured SPR usually being smaller and measured NU usually being larger than the simulated values. The measured values still demonstrate the advantages of the modified blended algorithm over the conventional blended algorithm, even though these differences are more difficult to see. In terms of SPR, for the tri-phase SLM the measured differences for the two types of blending are much less than predicted. The differences are much more pronounced for the quad-phase SLM. The situation is reversed for NU. There appears to be a floor to NU of 14%, and so for the quad-phase SLM, which is predicted to produce more uniform patterns, NU is only slightly different between the two blended algorithms. However, for the tri-phase SLM, NU is much larger for PRE and MDE, and this increase in NU is clearly seen in the measurements.

Since the experimental and simulated measurements of performance differ, it could be the case that the optimal performance occurs for different values of γ. This is explored for the case of mMD-PRE on a quad-phase phase-only SLM (Fig. 13). The performance measurements are compared with the simulated results (originally plotted in Fig. 11). Figure 13 shows that the measured diffraction efficiency is somewhat lower than the simulated, the measured NU is higher than the simulated, and the measured SPR is usually lower than the simulated. The shape of each measured curve is quite similar, which suggests that for our experimental SLM the simulated value of γ^* will be reasonably close to the optimal value of γ for experimental settings. While much closer agreement between measurement and theory has been demonstrated with fixed-pattern diffractive optics,[31] we believe that these results are in quite close agreement for programmable SLM's, which suffer from inaccuracies in setting the SLM phase identically on each pixel.[28] Also, interference effects that are due to multiple reflections from the SLM layers and other optical surfaces in the optical system can influence the measurements, especially in the case of NU measurements.[31]

One other possible source of discrepancy between theory and experiment for the NU measurements is the nonuniformity introduced by the frequency roll-off that is due to the subapertures of the SLM pixels. Since the desired portion of the diffraction pattern is along a diagonal, the intensity roll-off is proportional to $\mathrm{sinc}^4(x)$. We find that the closest correspondence between the simulated

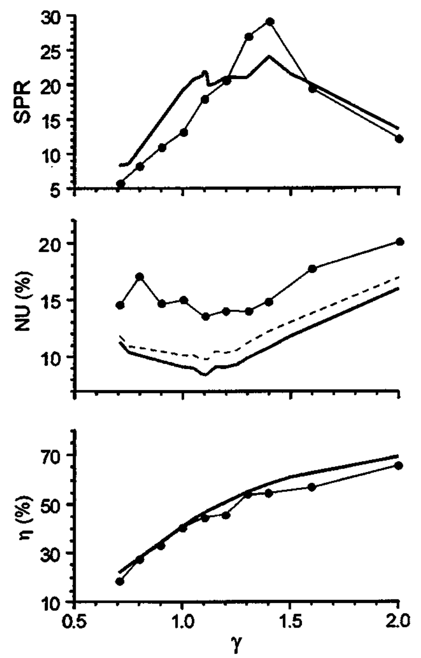


Fig. 13. Comparison between theory (thick solid curve) and experiment (thin solid curve and dots) of the performance of mMD-PRE on quad-phase phase-only SLM's. The dashed curve on the plots of NU shows the theory with the frequency response roll-off (which is to the aperture of the SLM pixels) taken into account.

and the experimentally measured NU occurs if a square pixel aperture of 56% fill factor is used to simulate the pixel-induced rolloff of the originally simulated diffraction pattern. For the particular design considered here, the intensity at the spot location furthest from the optical axis [i.e., the (7, 7) position] is reduced by 16% from the intensity at the spot that is closest to the optical axis. We find that recalculating NU with the additional roll-off would have increased NU by only 0.5%–1.5% over the results reported in Tables 1 and 2 for phase-only and quantized phase-only SLM's. Similarly, recalculated values of NU are plotted in Fig. 13 (dashed curve).[32] While theory and experiment are brought into closer agreement, the other factors considered above still appear to be the dominant sources of error.

6. CONCLUSIONS

A. Summary of Results

We have described and compared two possible ways of combining minimum-distance encoding (MDE) with pseudorandom encoding (PRE). The new modified approach maps the desired value to the closest value that can be achieved by pseudorandom encoding. Simulations with four types of coarsely quantized SLM characteristics clearly show that the modified blended algorithm mMD-PRE outperforms the conventional algorithm MD-PRE in fidelity as measured by peak levels of background noise across the full SBWP of the SLM and by the uniformity of the desired spot array.

Blending by either approach leads to significant improvements in performance over that originally reported in Ref. 24 for PRE and MDE used individually. Especially significant is the 2×–3× increase in diffraction efficiency over that with PRE alone as a result of blending. It is even possible to increase diffraction efficiency over the values reported for best fidelity (i.e., the result for γ^*) by trading off uniformity and SPR as controlled by the scaling parameter γ.

The experimentally implemented designs always showed that the modified blending outperformed the conventional blending, though the differences are not always as evident in all cases studied. This is attributed in large part to errors in controlling the phase of each SLM pixel identically. However, the measured diffraction patterns match the simulated diffraction patterns much more closely than was previously possible by using an optically addressed SLM in Ref. 22. The earlier SLM produced undesired noise orders because of its nonlinear properties and increased nonuniformity of the spot arrays because of its limited resolution. With these limitations absent, the spot arrays are more uniform and the background noise orders are primarily associated with the encoding algorithms and SLM quantization. A further desirable improvement in SLM's would be reduction of the on-axis spot that is due to reflections from the cover and the interfaces of the SLM and that is accentuated by the low-efficiency reflectance of modulated light. Even though further improvements in SLM's are desirable, these experimental results do demonstrate that the encoding algorithms proposed here perform in a manner quite similar to the simulations, thus making the algorithms suitable for use in real-time systems.

B. Implications for Future Research

While blended algorithms tend to improve the optical performance of SLM-based systems over the unblended PRE algorithms, there is additional overhead. Specifically, a search is required to find the optimal scaling parameter γ^*. At this time the only known way to perform this search numerically involves repeated fast-Fourier-transform-based simulations. Additional studies on encoding various functions could possibly lead to the development of a knowledge base that would provide a good *a priori* estimate of the optimal value γ^*. Alternatively, it may be possible to develop theoretical models of the performance of the encoding algorithms as a function of γ. A third possibility would be the inclusion in the optical system of an image sensor that records the far-field pattern and evaluates the performance on line. This would permit much faster evaluation and, as Fig. 13 illustrates, the *in situ* measurements could be used to compensate for the nonideal behavior and other vagaries of current SLM's.

In Section 4 we briefly considered trading off fidelity to increase diffraction efficiency. We achieved this by increasing the value of the blending parameter to increase the amount of MDE in the mMD-PRE algorithm. Additional trade-offs that favor diffraction efficiency can be envisioned by blending the conventional and modified MD-PRE algorithms. The proposed blending could be geometrically interpreted (see Fig. 1) as a mapping from the desired value $\mathbf{a}_c$ to a point on the exterior of the PRE (striped) region. The mapping can be considered a linear combination of the modified and conventional minimum-distance mappings. The actual implementation could be performed in at least two ways: (1) The value/point that

$\mathbf{a}_c$ is mapped to is pseudorandom encoded. (2) The modified and conventional encodings are randomly selected so that the value that $\mathbf{a}_c$ is mapped to is realized on average. Further analysis is required to determine if these approaches actually provide a second trade-off parameter in addition to γ or if one or both of these are alternative interpretations of mMD-PRE. Certainly, such studies may prove valuable, since Figs. 11 and 12 show that for the same value of γ the diffraction efficiency for MD-PRE is as much as 0.15 greater than the efficiency for mMD-PRE, especially when diffraction efficiency has a much higher premium than fidelity.

In conclusion, the performance of Fourier transform holograms from coarsely quantized SLM's can be significantly improved over minimum-distance, pseudorandom, and conventional blended encoding by instead using algorithms that blend pseudorandom encoding with modified minimum-distance encoding. While the new blended algorithm does not outperform blended algorithms for continuous-value phase-only SLM's, it may well be adequate to use coarse quantized SLM's in place of continuous SLM's in a number of applications. These algorithms may be especially useful for SLM developers because they permit early testing and evaluation with prototype devices that have greatly simplified and much less costly electrical addressing circuitry.

ACKNOWLEDGMENTS

This study was supported by Office of Naval Research grant N00014-96-1-1296, NASA cooperative agreement NCC5-222, and the Kentucky Space Grant Consortium.

Address correspondence to Robert W. Cohn at the location on the title page or by phone, 502-852-7077; fax, 502-852-1577; or e-mail, rwcohn@louisville.edu.

REFERENCES AND NOTES

1. N. C. Gallagher and B. Liu, "Method for computing kinoforms that reduces image reconstruction error," Appl. Opt. **12**, 2328–2335 (1973).
2. H. Stark, W. C. Catino, and J. L. LoCicero, "Design of phase gratings by generalized projections," J. Opt. Soc. Am. A **8**, 566–571 (1991).
3. M. P. Dames, R. J. Dowling, P. McKee, and D. Wood, "Efficient optical elements to generate intensity weighted spot arrays: design and fabrication," Appl. Opt. **30**, 2685–2691 (1991).
4. J. Bengtsson, "Kinoform design with an optimal-rotation-angle method," Appl. Opt. **33**, 6879–6884 (1994).
5. E. G. Johnson and M. A. Abushagur, "Microgenetic-algorithm optimization methods applied to dielectric gratings," J. Opt. Soc. Am. A **12**, 1152–1160 (1995).
6. J. N. Mait, "Understanding diffractive optic design in the scalar domain," J. Opt. Soc. Am. A **12**, 2145–2158 (1995).
7. J. N. Mait, "Diffractive beauty," Opt. Photon. News **9**, 21–25 (November 1998).
8. R. W. Cohn and M. Liang, "Pseudorandom phase-only encoding of real-time spatial light modulators," Appl. Opt. **35**, 2488–2498 (1996).
9. R. W. Cohn, A. A. Vasiliev, W. Liu, and D. L. Hill, "Fully complex diffractive optics via patterned diffuser arrays," J. Opt. Soc. Am. A **14**, 1110–1123 (1997).
10. B. R. Brown and A. W. Lohmann, "Complex spatial filter," Appl. Opt. **5**, 967–969 (1966).
11. W.-H. Lee, "Computer-generated holograms: techniques and applications," in *Progress in Optics*, E. Wolf, ed. (Elsevier, Amsterdam, 1978), Vol. 16, pp. 119–231.
12. W. J. Dallas, "Computer-generated holograms," in *The Computer in Optical Research*, B. R. Frieden, ed. (Springer, Berlin, 1980), Chap. 6, pp. 291–366.
13. R. W. Cohn and L. G. Hassebrook, "Representations of fully complex functions on real-time spatial light modulators," in *Optical Information Processing*, F. T. S. Yu and S. Jutamulia, eds. (Cambridge U. Press, Cambridge, UK, 1998), Chap. 15, pp. 396–432.
14. R. W. Cohn and W. Liu, "Pseudorandom encoding of fully complex modulation to bi-amplitude phase modulators," in *Diffractive Optics and Micro-optics*, Vol. 5 of 1996 OSA Technical Digest Series (Optical Society of America, Washington, D.C., 1996), pp. 237–240.
15. L. G. Hassebrook, M. E. Lhamon, R. C. Daley, R. W. Cohn, and M. Liang, "Random phase encoding of composite fully complex filters," Opt. Lett. **21**, 272–274 (1996).
16. R. W. Cohn and M. Liang, "Approximating fully complex spatial modulation with pseudorandom phase-only modulation," Appl. Opt. **33**, 4406–4415 (1994).
17. R. D. Juday, "Optimal realizable filters and the minimum Euclidean distance principle," Appl. Opt. **32**, 5100–5111 (1993).
18. R. W. Cohn, "Pseudorandom encoding of fully complex functions onto amplitude coupled phase modulators," J. Opt. Soc. Am. A **15**, 868–883 (1998).
19. J. P. Kirk and A. L. Jones, "Phase-only complex-valued spatial filter," J. Opt. Soc. Am. **61**, 1023–1028 (1971).
20. L. B. Lesem, P. M. Hirsch, and J. A. Jordon, Jr., "The kinoform: a new wavefront reconstruction device," IBM J. Res. Dev. **13**, 150–155 (1969).
21. J. L. Horner and P. D. Gianino, "Phase-only matched filtering," Appl. Opt. **23**, 812–816 (1984).
22. R. W. Cohn and M. Duelli, "Ternary pseudorandom encoding of Fourier transform holograms," J. Opt. Soc. Am. A **16**, 71–84 (1999).
23. M. W. Farn and J. W. Goodman, "Optimal maximum correlation filter for arbitrarily constrained devices," Appl. Opt. **28**, 3362–3366 (1989).
24. R. D. Juday, "Correlation with a spatial light modulator having phase and amplitude cross coupling," Appl. Opt. **28**, 4865–4869 (1989).
25. M. Montes-Usategui, J. Campos, and I. Juvells, "Computation of arbitrarily constrained synthetic discriminant functions," Appl. Opt. **34**, 3904–3914 (1995).
26. R. D. Juday and J. Knopp, "HOLOMED—an algorithm for computer generated holograms," in *Optical Pattern Recognition VII*, D. P. Casasent and T. Chao, eds., Proc. SPIE **2752**, 162–172 (1996).
27. R. W. Cohn, "Analyzing the encoding range of amplitude-phase coupled spatial light modulators," Opt. Eng. **38**, 361–367 (1999).
28. U. Krackhardt, J. N. Mait, and N. Streibl, "Upper bound on the diffraction efficiency of phase-only fanout elements," Appl. Opt. **31**, 27–37 (1992).
29. D. J. Cho, S. T. Thurman, J. T. Donner, and G. M. Morris, "Characteristics of a 128 × 128 liquid crystal spatial light modulator for wave-front generation," Opt. Lett. **23**, 969–971 (1998).
30. A. Bergeron, J. Gauvin, F. Gagnon, D. Gingras, H. H. Arsenault, and M. Doucet, "Phase calibration and applications of a liquid-crystal spatial light modulator," Appl. Opt. **34**, 5133–5139 (1995).
31. M. Duelli, D. L. Hill, and R. W. Cohn, "Frequency swept measurements of coherent diffraction patterns," Appl. Opt. **37**, 8131–8133 (1998).
32. The influence that is due to roll-off may at first seem to be surprisingly small, but calculating the effect of this roll-off on the nonuniformity/standard deviation of the ideally uniform spot array gives NU = 4.2%. The small influence of the rolloff on NU is further explained by the fact that the simulated values of NU are generally greater than 4% and that standard deviations, rather than being additive, add as the square root of the sum of the squares.

Reductive Dehalogenation of Trichloroethylene with Zero-Valent Iron: Surface Profiling Microscopy and Rate Enhancement Studies

J. Gotpagar,[†] S. Lyuksyutov,[‡] R. Cohn,[‡] E. Grulke,[†] and D. Bhattacharyya*,[†]

Department of Chemical and Materials Engineering, University of Kentucky, Lexington, Kentucky 40506, and The Electrooptics Research Institute, University of Louisville, Louisville, Kentucky 40292

Received March 18, 1999. In Final Form: July 9, 1999

Mechanistic aspects of the reductive dehalogenation of trichloroethylene using zerovalent iron are studied with three different surface characterization techniques. These include scanning electron microscopy, surface profilometry, and atomic force microscopy. It was found that the pretreatment of an iron surface by chloride ions causes enhancement in the initial degradation rates. This enhancement was attributed to the increased roughness of the iron surface due to crevice corrosion obtained by pretreatment. The results indicate that the "fractional active site concentration" for the reactive sorption of trichloroethylene is related to the number of defects/abnormalities present on the surface of the iron. This was elucidated with the help of atomic force microscopy. Two possible mechanisms include (1) a direct hydrogenation in the presence of defects acting as catalyst and (2) an enhancement due to the two electrochemical cells operating in proximity to each other. The result of this study has potential for further research to achieve an increase in the reaction rates by surface modifications in a practical scenario.

1. Introduction

Since the original studies by Gillham and O'Hannesin[1,2] who proposed the use of reductive dehalogenation reaction for environmental remediation, several studies have been published which deal with the reaction of trichloroethylene (TCE) with zerovalent iron. Matheson and Tratnyek[3] were the first to report a detailed kinetic and mechanistic study of this reaction with several contaminants of concern. It is well-known now that though these reactions are faster compared to the natural biotic and abiotic processes, the rates are nonetheless too low to be feasible for ex situ applications. Therefore, the focus so far in this area is directed toward the in situ applications. Both in situ reactive barriers and above ground reactors have been developed for this purpose. Several test installations have already been completed at contaminated sites, and more are being planned.[4–7] However, little information is available on the exact mechanism of the reductive dehalogenation using zerovalent iron. The effective design and operation of systems involving zerovalent metals would be greatly improved by a more detailed, process-level understanding of the mechanism by which these contaminants degrade. Furthermore, improving the rates of these processes would also make the ex situ applications feasible and reduce the remediation time and would also be cost-effective. In this paper, an attempt is made to gain an understanding of the surface phenomena on the surface of iron during the reductive dehalogenation of TCE with the ultimate goal of increasing the rate of reaction. Atomic force microscopy (AFM), surface profilometry, and scanning electron microscopy (SEM) were used to analyze the surface features of the iron. In addition, the relation between the metal dissolution process occurring during dehalogenation of chlorinated organics to the classical crevice corrosion mechanism of iron in the presence of chloride ion is described. Only recently, promising use of AFM was suggested by Boronina et al.[8] for such environmental applications. In this study, AFM is also found to be important for the indication of crevice corrosion, as will be discussed later on.

2. Background. Role of the Metallic Surface

2.1. Role of the Metal Surface on Electron Transfer. Although considerable advancement has been made recently in identifying the product distribution,[9–11] to date the exact surface mechanism for TCE degradation by iron is not known. There is general agreement that electron transfer at the metal surface is required. This observation was used by Gotpagar et al.[12] and Boronina et al.[13] to develop the macroscopic model. Recent publications[3,5,13–18] have repeatedly emphasized the importance of the metal

* To whom correspondence should be addressed. Phone: (606) 257-2794. Fax: (606) 323-1929. E-mail: db@engr.uky.edu.

[†] University of Kentucky.

[‡] University of Louisville.

[1] Gillham, R. W.; O'Hannesin, S. F. Metal-catalyzed Abiotic Degradation of Halogenated Organic Compounds. Paper presented at the 1992 IAH Conference on Modern Trends in Hydrogeology, Hamilton, Ontario, Canada, May 10–13, 1992.

[2] Gillham, R. W.; O'Hannesin, S. F. *Groundwater* **1994**, *32*, 958.

[3] Matheson, L. J.; Tratnyek, P. G. *Environ. Sci. Technol.* **1994**, *28*, 2045.

[4] Gillham, R. W.; O'Hannesin, S. F.; Orth, W. S. Metal Enhanced Abiotic 5. Degradation of Halogenated Aliphatics: Laboratory Tests and Field Trials. Paper presented at the 1993 HazMat Central Conference, Chicago, IL, March 9–11, 1993.

[5] Gillham, R. W. *Prepr. Extended Abstr. Am. Chem. Soc.* **1995**, *35*, 691.

[6] Puls, R. W.; Powell, R. M. *Environ. Sci. Technol.* **1997**, *31*, 2244.

[7] Yamane, C. L.; Gallinatti, J. D.; Szerdy, F. S.; Delfino, T. A.; Hankins, D. A.; Vogan, J. L. *Prepr. Extended Abstr. Am. Chem. Soc.* **1995**, *35*, 792.

[8] Boronina, T. N.; Lagadic, I.; Sergeev, G. B.; Klabunde, K. J. *Environ. Sci. Technol.* **1998**, *32*, 2614.

[9] Roberts, A. L.; Wells, J. R.; Campbell, T. J.; Burris, D. R. *Environ. Toxicol. Chem.* **1997**, *16*, 625.

[10] Burris, D. R.; Delcomyn, C. A.; Smith, M. H.; Roberts, A. L. *Environ. Sci. Technol.* **1996**, *30*, 3047.

[11] Arnold, W. A.; Roberts, A. L. *Environ. Sci. Technol.* **1998**, *32*, 3017.

[12] Gotpagar, J.; Grulke, E.; Tsang, T.; Bhattacharyya, D. *Environ. Prog.* **1997**, *16*, 137.

[13] Boronina, T.; Klabunde, K. J.; Sergeev, G. *Environ. Sci. Technol* **1995**, *29*, 1511.

10.1021/la990325x CCC: $18.00
Published on Web 09/18/1999

surface area in the process. It was also found that the presence of carbonate- or oxide-forming species in the water leads to an inert layer of metal oxide or metal carbonate forming on the metal surface. This layer greatly reduces the overall reaction rate.[18-21] Thus, it is quite clear that the metallic surface is the controlling factor in TCE degradation.

Current research efforts are directed toward obtaining enhancements in reaction rates. Bimetallic complexes have been used to generate higher chlorinated organic degradation rates.[22-25] In particular, Li and Klabunde have shown that doping various zinc samples with palladium, silver, and nickel resulted in much higher pseudo-first-order degradation rates, with some systems showing as much as 150 times higher values. The standard oxidation–reduction potentials of these metals relative to a hydrogen electrode are +0.987 (Pd), +0.799 (Ag), −0.250 (Ni), −0.440 (Fe), and −0.763 (Zn). For each metal pair, the metal with the lower potential is preferentially dissolved, releasing electrons, e.g., the zinc dissolution in the Pd–Zn pair.

When iron is used, the chemistry at the metal surface during a dehalogenation can be similar to that of classical crevice corrosion of steel in the presence of chloride ions.[26] In the presence of oxygen, iron metal is oxidized, releasing electrons, which can be used in the reduction reaction of water plus oxygen, generating hydroxide ions. Depletion of oxygen in crevices leads to an excess of positive charges in the local solution, causing the diffusion of chloride ions into these spaces and increasing the metal dissolution. Insoluble metal hydroxides can form and coat the exterior surface, reducing the rate of metal dissolution.

The presence of a second metal electropositive relative to iron would provide surfaces that might exhibit less fouling and could continue to supply electrons to a complete oxidation–reduction process cycle. Furthermore, the presence of a second, nonreactive metal would accelerate the dissolution of the reactive metal in crevices. This surface area affect would add to the faster chloride ion migration within the crevices. For example, Li and Klabunde[25] showed that more porous zinc was formed during their degradations. In their case, the second electropositive metal could have accelerated the process both by providing a nonfouling surface for reduction reactions and by promoting crevice formation in the zinc-rich areas of the material (enhanced via the chloride ion diffusion process). Surface Fe^{2+} has been shown to play a very important role in the degradation of chlorinated organics.

The decrease in the reaction rate with time for zerovalent iron systems may be due to changes in the iron surface morphology. For example, the formation of insoluble iron hydroxides could foul the surface, reducing the reduction process and slowing the development of crevices. One approach reported in this area was the use of bimetallic complexes.[25] Matheson and Tratnyek[3] have argued that commonly used bimetallic systems include a hydrogenation catalyst such as Ni or Pd that can further enhance the rate. Although the results obtained with such bimetallic systems seem encouraging, the enhanced degradation rates have only been studied for short periods of time. The outer layer of these materials are quickly covered with iron oxides,[22,27,28] thus causing rate reduction.

Another new approach for iron surface (without the use of bimetallic systems) regeneration is the use of ultrasound. This is confirmed by a recently published study,[29] which observed overall rate enhancement by a factor of 40 for dechlorination of CCl_4 by zerovalent iron, in the presence of ultrasound waves. The reason for this enhancement was attributed to the continuous cleaning and activation of the Fe^0 surface by ultrasound waves, and the enhanced rates of mass transport resulting from cavitation. Characterization of the iron surface can thus provide helpful insight into the mechanistic aspects of the reaction, thereby potentially leading to the enhancement of process effectiveness.

2.2. Sorption of TCE onto the Iron Surface. The first step in the reaction of TCE reduction with Fe^0 is the TCE sorption onto the iron surface. The sorption of TCE takes place at two different sites—reactive sites and nonreactive sites. Moreover, Burris et al.[30] have shown that the majority of this sorption takes place at the nonreactive sites. It was shown that the adsorption of TCE to iron follows the modified Langmuir type isotherm given by

$$C^{\mathrm{S}}_{\mathrm{TCE}} = kb(C^{\mathrm{W}}_{\mathrm{TCE}})^{M}/(1 + k(C^{\mathrm{W}}_{\mathrm{TCE}})^{M}) \qquad (1)$$

where $C^{\mathrm{S}}_{\mathrm{TCE}}$ and $C^{\mathrm{W}}_{\mathrm{TCE}}$ are the sorbed concentration (nmol/g) and aqueous phase concentration (nmol/ml), respectively. M, k, and b are generalized Langmuir coefficients.[30] In our earlier published study,[17] the concept of fractional active site concentration was introduced to take into account this difference in the sorption behavior. The fractional active site concentration was defined as

$$A_{\mathrm{S}} = C^{\mathrm{S}*}_{\mathrm{TCE}}/C^{\mathrm{S}}_{\mathrm{TCE}} \qquad (2)$$

where A_{S} is the fractional active site concentration and $C^{\mathrm{S}*}_{\mathrm{TCE}}$ is the concentration of TCE on reactive sites (nmol/g). It was shown that, with this concept of fractional active site concentration, the model developed was able to predict the degradation of TCE with time quite accurately over the entire course of time. The equation indicating the decline in the concentration of TCE in a simple two-phase

[14] Lipczynska-Kochany, E.; Harms, S.; Milburn, R.; Sprah, G.; Nadarajah, N. *Chemosphere* **1994**, *29*, 1477.

[15] Helland, B. R.; Alvarez, P. J. J.; Schnoor, J. L. *J. Haz. Mater.* **1995**, *41*, 205.

[16] Warren, K. D.; Arnold, R. L.; Bishop, T. L.; Lindholm, L. C.; Betterton, E. A. *J. Haz. Mater.* **1995**, *41*, 217.

[17] Gotpagar, J.; Grulke, E.; Bhattacharyya, D. *J. Haz. Mater.* **1998**, *62*, 243.

[18] MacKenzie, P. D.; Baghel, S. S.; Eykholt, G. R.; Horney, D. P.; Salvo, J. J.; Sivavec, T. M. American Chemical Society Extended Abstract, Industrial and Engineering Chemistry Division, Sept 17–20, 1995, pp 59–62.

[19] Agrawal, A.; Liang, L.; Tratnyek, P. G. American Chemical Society Extended Abstract, Industrial and Engineering Chemistry Division, Sept 17–20, 1995, p 54.

[20] Liang, L.; Goodlaxson, J. D. American Chemical Society Extended Abstract, Industrial and Engineering Chemistry Division, Sept 17–20, 1995, pp 46–49.

[21] Sivavec, T. M.; Horney, D. P.; Baghel, S. S. American Chemical Society Extended Abstract, Industrial and Engineering Chemistry Division, Sept 17–20, 1995, pp 42–45.

[22] Muftikian, R.; Fernando, Q.; Korte, N. *Water Res.* **1995**, *29*, 2434.

[23] Roberts, A. L.; Fennelly, J. P. *Environ. Sci. Technol.* **1988**, *32*, 1980.

[24] Korte, N. E.; Grittini, C.; Muftikian, R.; Fernando, Q.; Liang, L.; Clausen, J. American Chemical Society Extended Abstract, Industrial and Engineering Chemistry Division, Sept 17–20, 1995, pp 51–53.

[25] Li., W.; Klabunde, K. J. *Croat. Chem. Acta* **1998**, *71*, 853.

[26] Fontana, M. G.; Greene, N. D. *Corrosion Engineering*, 2nd ed., McGraw-Hill Book Co.: New York, 1978.

[27] Muftikian, R.; Nebesney, K.; Fernando, Q.; Korte, N. *Environ. Sci. Technol.* **1996**, *30*, 3593.

[28] Sivavec, T. M.; Mackenzie, P. D.; Horney, D. P. American Chemical Society Extended Abstract, Industrial and Engineering Chemistry Division, April 13–17, 1997, pp 83–85.

[29] Hung, H.-M.; Hoffmann, M. R. *Environ. Sci. Technol.* **1998**, *32*, 3011.

[30] Burris, D. R.; Campbell, T. J.; Manoranjan, V. S. *Environ. Sci. Technol.* **1995**, *29*, 2850.

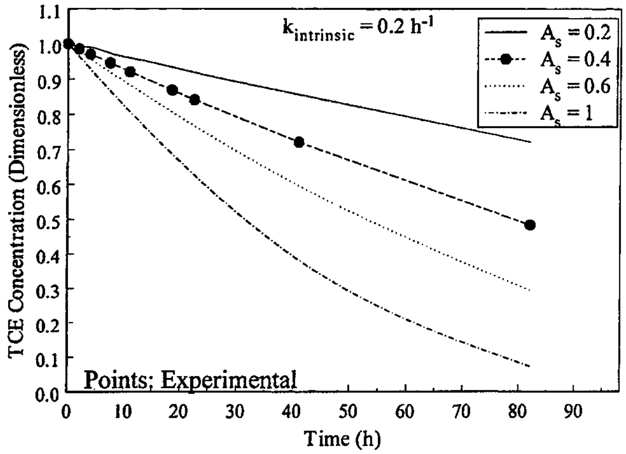


Figure 1. Effect of fractional active site concentration on the simulated TCE degradation profile with zerovalent iron.

(solid–liquid) closed system with iron and TCE dissolved in water was obtained as

$$\frac{dC^{W}_{TCE}}{dt} = -\frac{(kb(C^{W}_{TCE})^{M}/(1 + k(C^{W}_{TCE})^{M}))k_{\text{intrinsic}}A_{S}m_{Fe}}{V^{W}[1 + (m_{Fe}Mkb(C^{W}_{TCE})^{M-1})/(V_{W}(1 + k(C^{W}_{TCE})^{M})^{2})]} \tag{3}$$

where m_{Fe} is the amount of iron used in the reaction system (g), V_W is the volume of the aqueous phase used (L), and $k_{\text{intrinsic}}$ is the intrinsic value of the degradation rate constant (h^{-1}). The above equation is also consistent with the observed zero-order reaction behavior at higher TCE concentrations. The result of this modeling analysis[31] indicated that the actual intrinsic degradation constant and the observed value of the same are related by

$$k_{\text{obs}} = (kbA_{S}m_{Fe}/3V_{W})k_{\text{intrinsic}} \tag{4}$$

where k_{obs} is the observed value of the degradation constant (h^{-1}). It is instructive to note from the above equation that even if the actual intrinsic degradation constant value for the reaction of reductive dehalogenation is high, due to the small coverage of the reactive sites, the overall degradation constant value would be lowered to a great extent (almost an order of magnitude). This is also clear from Figure 1, which shows the effect of fractional active site concentration on the calculated TCE degradation. As expected, increasing the value of A_S causes an increase in the degradation rate. The calculated curve for $A_S = 0.4$ fits the experimental data quite well (with $k = 0.0207$, $b = 882$, and $M = 0.655$, values obtained from Burris et al.[30]). Thus, one should look at different ways to improve the active surface area to explore the possible enhancements in the degradation rates. The role of surface area was further verificated by Li and Klabunde,[25] in which ultrafine zinc coated with small amounts of palladium gave the highest reactivity for carbon tetrachloride dechlorination. As mentioned in the previous section, the oxide layer present on the surface of iron was found to slow the TCE degradation. In other words, the active surface available for the TCE degradation is reduced as time progresses. Continuous cleaning of the surface (increasing A_S) was found to increase the degradation rate considerably.[29]

It is clear that the low rates of dehalogenation obtained with zerovalent iron are partly due to the abundance of the nonreactive sorption and also due to the presence of the oxide layer acting as an additional barrier for electron transfer. It is not yet clear what exactly are the reactive sites. We hypothesized that the defects/abnormalities present on the surface of iron contribute to the reactive sites. The basis for this hypothesis lies in the corrosion literature. Review of the same indicated that these defects cause increased dissolution of the metal in the corrosion process, due to the phenomenon called localized corrosion.

Chloride pretreatment was used to increase the number of defects on the iron surface. This has two advantages. First, it is widely accepted in the literature that pitting is normally initiated by the aggressive anions such as halide ions. Since pitting corrosion is greatly enhanced by chloride ions,[32] it is possible that the dechlorination might favor further degradation if enough chloride accumulates. Thus, the reaction can act as an autocatalytic process. Second, in the presence of halide ions, the passive oxide layers formed on the surface of iron are known to break apart. At least two studies so far have indicated the presence of such an effect.[15,33] However, due to very small concentrations of the chloride ions present in the solutions, such autocatalytic effects are not evident in the reaction times studied so far. Moreover, though the driving force for reaction is corrosion of the metal, increasing the corrosion did not necessarily increase the degradation rate due to the competing water dissociation reaction.[12] Therefore, the corrosion process that was important for the dechlorination process was thought to be different in nature.

In this paper, we explore the possibility of induced pitting on the surface of iron as one of the techniques to improve the rates of the degradation process. The phenomenon of pitting corrosion is also seen in the scanning electron micrographs of iron samples observed over longer reaction times. To corroborate this further, iron was pretreated with chloride ions to introduce defects on the surface, and its effect on the TCE degradation rate was analyzed. These defects/pits present on the iron surface were found to be the controlling factor in determining the rate of reductive dehalogenation. We find that increasing the number of these surface abnormalities increases the rates considerably. Thus, the fractional active site concentration, as given by eq 2, is attributed to the number of defects/pits present on the surface of iron. The following section outlines the detailed experimental procedure adopted for the studies.

3. Experimental Section

3.1. Pretreatment of Iron. The electrolytic iron obtained from Fisher Scientific (100 mesh, 150 μm) was first treated with 1 M NaCl solution. Before treatment, the brine solution was heated to 100 °C, as the enhanced pitting is reported at higher temperatures.[34] Various other factors such as chloride concentration, pH of the solution, etc. influence the morphology of pits formed. Lower pH values have been found to give consistently higher pit formation even in the presence of small chloride concentrations. In the current approach, no attempt was made to treat the iron at lower pH, since at such low values loss of iron through dissolution would also be increased. To increase pit

[31] Gotpagar, J. Reductive Dehalogenation of Trichloroethylene (TCE) with Zerovalent Iron: Reaction Mechanisms and Transport Modeling. Ph.D. Dissertation, University of Kentucky, 1998.

[32] Bardwell, J. A.; Fraser, J. W.; MacDugall, B.; Graham, M. J. *J. Electrochem. Soc.* **1992**, *139*, 366.

[33] Johnson, T. L.; Fish, W.; Gorby, Y. A.; Tratnyek, P. G. *J. Contam. Hydrol.* **1998**, *29*, 379.

[34] Sato, N. *Corrosion* **1989**, *45*, 354.

formation, higher chloride concentration (1 M) was used and the iron samples were treated for 5 days.

The iron samples were washed with deoxygenated, deionized water for 3–4 cycles to remove the traces of chloride on the surface. This is a conservative approach because the presence of this species on the surface would have a benign effect (autocatalytic) if any. The samples were then immediately soaked in TCE solutions of known concentration, and placed on the rotary shaker at 5 rpm. At selected times, the aqueous samples were analyzed for TCE. The Hewlett-Packard 5890 Series II gas chromatograph, with an attached MS 5971A quadruple mass detector, was used for TCE analysis using a fused capillary column, J&W Scientific, DB-624. The analytical method followed the EPA method 624, with the following temperature program: oven temperature of 35 °C (4 min) to 200 °C at 6 °C/min, hold at 200 °C for 4 min. The carrier gas was zero-grade (high-purity) helium with a flow rate of 7.5 mL/min, and the MS scan range was $m/z = 35-260$ at 0.6 s/scan.

The chloride-treated iron samples were further characterized using SEM, surface profilometry, and AFM techniques. Initial AFM scans on Fisher electrolytic iron (100 mesh) showed that the AFM tip always moves the iron grains. To prevent this, the iron granules were embedded in a matrix of polyethylene (LDPE). This approach was also unsuccessful because of movement of grains. Therefore, iron chips were used in a separate experiment to be analyzable by AFM. The chips (35 × 10 × 3 mm) were obtained from hot rolled steel manufactured by Harbor Steel Corp., Lexington, KY. The experimental procedure followed for degradation studies was the same as that with Fisher iron.

3.2. Scanning Electron Microscopy. SEM provides an effective method for the characterization of the surfaces of samples. A narrow beam of electrons with kinetic energy in the range of 0–25 kV is incident on the iron surface. The sample was glued to the sample holder with colloidal graphite. The scanning electron microscope was a S-2300 model from Hitachi, and the images were recorded with 3000× magnification.

3.3. Surface Profilometric Studies. The surface profilometry studies were performed using a WYKO NT-2000 scanning white light interoferometry profiler. The WYKO is a noncontact optical device capable of measuring surface heights between 4 Å and 1 mm. Smooth surfaces can be measured in the phase-shifting interferometry (PSI) mode, while vertical-scanning interferometry (VSI) allows the measurement of rough surfaces and steps, without resorting to phase-unwrapping algorithms. In our studies we used the VSI mode to study the surfaces of treated/untreated iron.

3.4. Atomic Force Microscopy. A Park Autoprobe M5 atomic force microscope was used to profile the surface of treated/untreated samples of the iron chips before and after the chloride treatment. The atomic force microscope probes the surface of a sample with a sharp tip three μm long and 20 nm in diameter located at the free end of a cantilever. Forces between the tip and the sample surface cause the cantilever to bend and deflect. A laser beam reflected from the cantilever hits a detector area. The detector measures the current, proportional to the cantilever deflection, as the tip is scanned over the sample. These current measurements allow a computer to generate an image of surface topography. There are two main modes in operating an atomic force microscope: contact mode and noncontact mode. In the noncontact mode, the cantilever vibrates on the order of tens to hundreds of angstroms above the sample surface, and the interatomic force between the cantilever and sample is attractive. In the contact mode, the cantilever is held a few angstroms above the sample surface, and the interatomic force between the cantilever and the sample is repulsive. Since available iron chips have a homogeneous surface, we selected the contact mode for our studies.

The silicon tips attached to a cantilever with a low spring constant were used. The magnitude of the net force exerted to the sample varied from 8 to 15 nN (nanoNewtons). The scan speed was 100 mm/s. Typically a 100 mm × 100 mm field is scanned. Two morphologies were observed on the surface of untreated iron chips. A shiny (highly reflective) area represents the layers of the iron oxide. Dark areas (weak reflectivity) are associated with naturally occurring rust. On treated iron chips, a third morphology was observed. This was a grainy structure of iron that becomes apparent after the layer of iron oxide is removed.

4. Results and Discussion

4.1. Localized/Pitting Corrosion. Pitting is normally initiated by the aggressive anions such as halide ions. Several studies exist in the literature which talk about the mechanism of breakdown of passive films present on the surface of iron in the presence of chloride-containing media.[35,36] The breakdown requires a minimum amount of electrocapillary energy, whatever the mechanism of breakdown. Furthermore, the kinetic data of Pou et al.[36] reveal that the breakdown of oxide layers was consistent with the ion exchange processes, point-defect models, and hydrated polymeric oxide model. For a review of the mechanistic aspects of the breakdown of oxide layers, the reader is referred to the above two references. In this paper, only results pertaining to our studies will be explained.

At this point, it is instructive to examine the SEM photographs of the iron (Fisher iron) samples obtained after TCE degradation. Parts a and b of Figure 2 show the scanning electron micrographs of the fresh iron surface and the iron surface after 81 h of degradation of TCE, respectively. The effect of localized corrosion is not very evident from these photographs. Figure 2c shows SEM of the iron surface after 120 days of reaction time. It can be seen that the localized corrosion is much more pronounced after longer reaction times. This indicates the breakdown of the precipitates with chloride ions. This is also consistent with the result of Helland et al.,[15] who reported a 60% increase in the CCl_4 dechlorination rates with increased contact time, for the specific case of batch systems with zerovalent iron. Thus, this can be interpreted as an autocatalytic effect observed due to pitting corrosion in the presence of chloride ions generated as a product of reaction. As can be seen from Figure 2c, the corrosion indeed appears to be localized and can be termed as crevice corrosion.

It is quite clear that chloride ions are responsible for the crevice corrosion observed. Therefore, we hypothesize that increasing this form of corrosion should increase the degradation rates. To investigate this, we deliberately treated the iron surface with chloride ions prior to its exposure to TCE solution. The morphological changes expected due to the attack of chloride ions are defects on the surface with trenches and peaks. The effect of this surface modification on the TCE degradation has been outlined in the subsequent sections.

4.2. Effect of Surface Pretreatment on TCE Degradation. Figure 3 shows the results of TCE degradation profiles obtained after the chloride treatment of Fisher electrolytic iron (100 mesh). The results obtained with untreated iron are also compared in the figure. These results show that TCE degradation is indeed increased by the chloride treatment. This increase is most prominent in the reaction at early times. At later times, the effect of chloride pretreatment provides little improvement (not shown). Figure 3 shows the pronounced effect of chloride treatment on the TCE degradation. This is also evident from Figure 4, where the pseudo-first-order rate constants for the degradation are plotted on the basis of the initial rates. As can be seen from the figure, chloride pretreatment provided almost 2-fold increase (from 0.019 to 0.037 h^{-1}) in the initial rate constant. Similar enhancements in the

[35] Sato, N. *J. Electrochem. Soc.* **1982**, *129*, 255.

[36] Pou, T. E.; Murphy, O. J.; Young, V.; Bockris, J. *J. Electrochem. Soc.* **1984**, *131*, 1243.

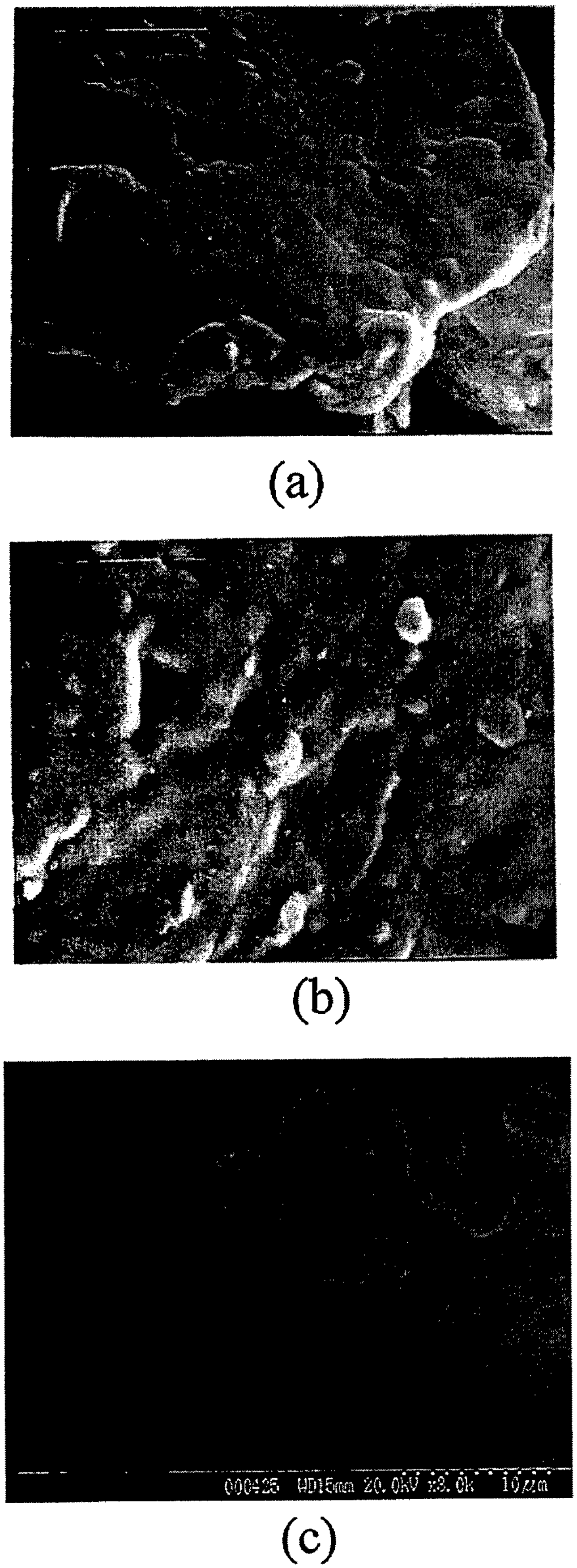


Figure 2. Evidence of crevice corrosion. SEM photographs of the iron surface: (a) fresh iron surface; (b) iron surface after 81 h of reaction; (c) iron surface after 120 days of reaction time.

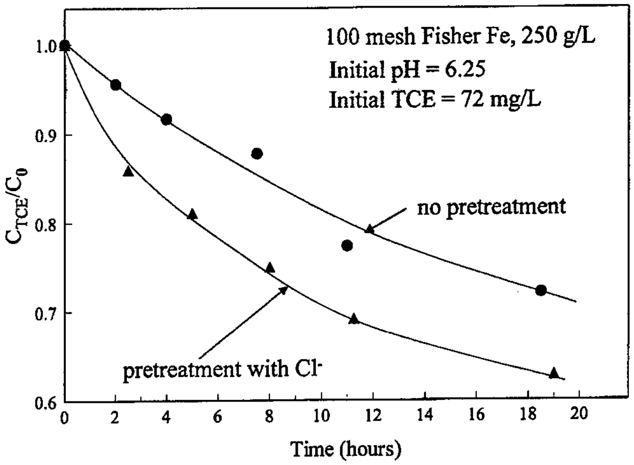


Figure 3. Effect of chloride treatment on the TCE degradation: (a) overall degradation profile; (b) enhancement in the initial degradation rates.

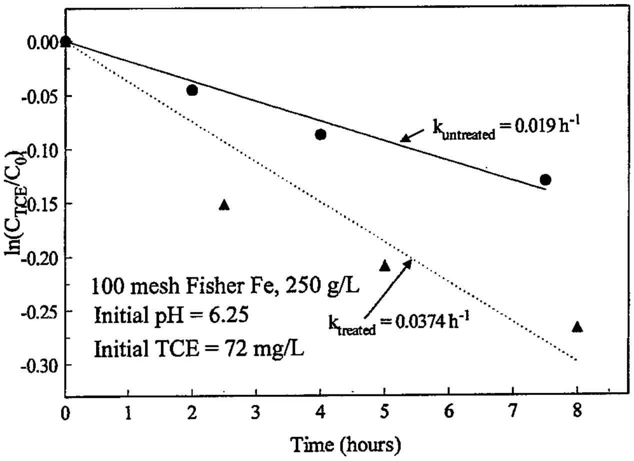


Figure 4. Effect of chloride treatment on the initial TCE degradation rate constant.

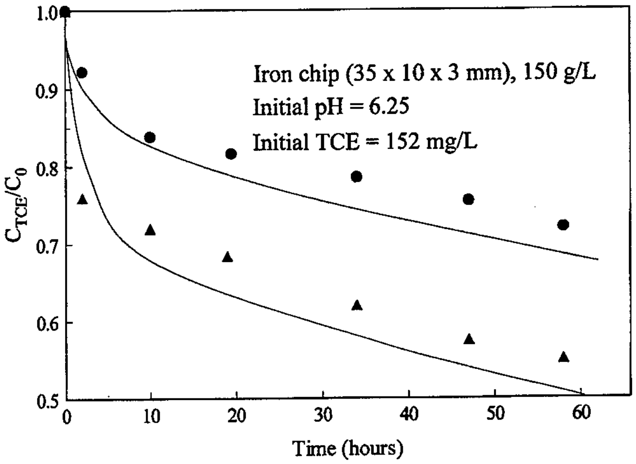


Figure 5. Effect of chloride treatment on TCE degradation with iron chips.

TCE degradation rates are also observed for iron chips from chloride pretreatment (Figure 5). As can be seen from Figure 5, the initial rates were again enhanced by the chloride treatment. It should be noted that the iron chips used in this study have very low external surface area, resulting in the low degradation rates. However, even with such low surface areas, the effect of chloride treatment is quite evident. Comparison of Figures 3 and

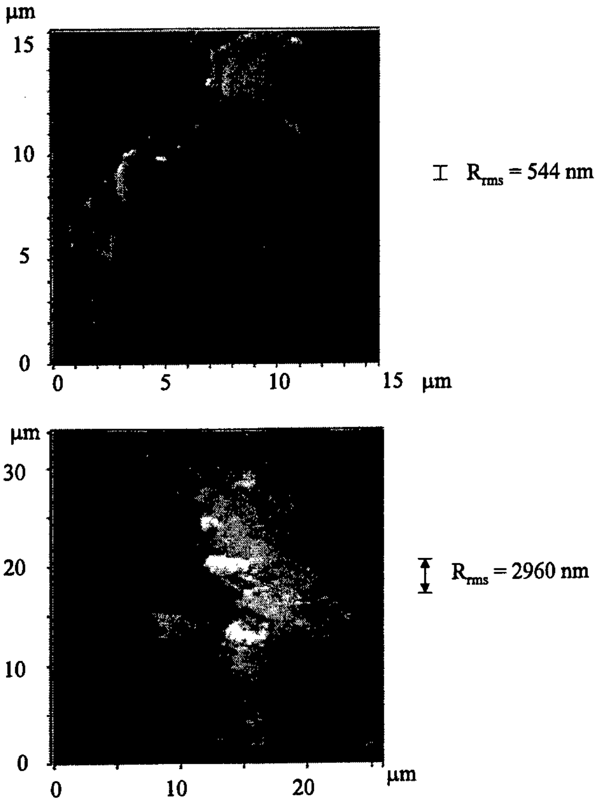


Figure 6. Surface profilometric data on the iron surface: (a, top) fresh iron surface; (b, bottom) iron surface after chloride treatment.

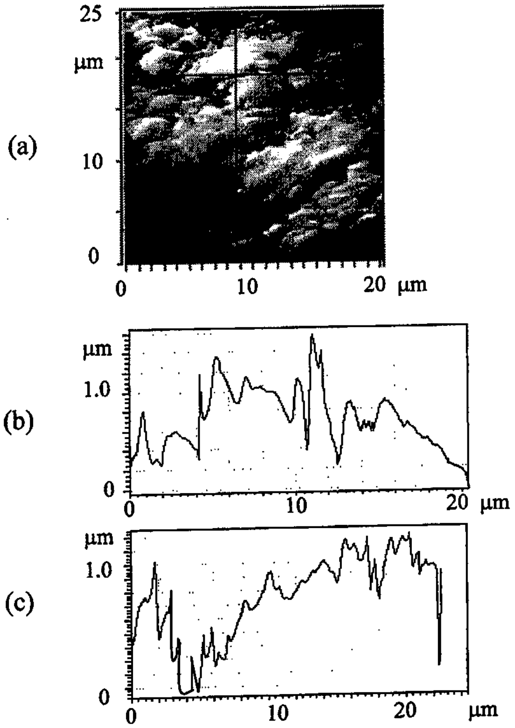


Figure 7. Surface profilometric data on the iron surface after 5 days of chloride pretreatment: (a) surface image; (b) vertical profile of surface roughness; (c) horizontal profile of surface roughness.

5 also reveals that the initial decline obtained with iron chips is much faster than that with 100 mesh Fisher iron fillings. The reason for this is not clear, although it is possible that even though iron chips have lower external surface areas, the additional internal surface area created by crack formation after the pretreatment might be more than that for the 100 mesh iron filings.

4.3. Surface Profilometric Results. Gray-scale renditions (obtained using the WYKO) of the surface topography for the fresh iron and chloride-treated iron are shown in Figure 6. These images show that the chloride-treated surface is much more rough than the untreated surface. This can be quantified by calculating the standard deviation of the 2D surface profile. The root-mean-square roughness for the untreated sample is 544 nm, and that for the treated sample is 2960 nm, which is 5.4 times greater than that of the untreated sample. This is also evident from the height variation along the x and y coordinate directions, as depicted in Figure 7. The fresh iron surface, i.e., covered with oxide, does have some variation (not shown) or, in other words, defects on the surface. We hypothesized that these defects are indeed the places where reaction takes place. In other words, the active sites[17,30] where the reactive sorption is assumed to occur are these defects or cracks on the surface. Attempts to increase this number of cracks by chloride pretreatment indeed resulted in the increase in the reaction. This further corroborates our hypothesis, which is also clear from Figure 7. To check whether increased roughness is actually due to the presence of chloride, iron was also treated in deionized water for 5 days in the absence of chloride. Comparison of the surface plot of corroded iron in the presence and absence of chloride (not shown) revealed that the roughness observed on the surface indeed was due to the pitting effect produced by chloride treatment.

The effect of this surface treatment is reflected in the increase in the degradation rates, which is shown in Figures 3–5. Thus, our hypothesis that such a surface roughness is beneficial for TCE degradation seems to be justified. Furthermore, the decline in the enhancement at longer times can be explained with the help of Figure 8. Examination of Figure 8 shows that though the iron surface after reaction appears rough at first glance, the height variations (peaks/valleys) on the surface have diminished compared to those on the freshly treated iron surface. As a result, there is no more rate enhancement because of reduction in the number of active sites (defects/pits) available as hypothesized earlier. The average roughness was found to be considerably less for the sample after 60 h of reaction, indicating that pits initially present are either being filled over the course of reaction or are being eaten away due to reaction, leaving a smoother metal surface as found by Boronina et al.[8] in the case of zinc. The possible reasons for this again can be formation of precipitates. Carrying out the reaction in chloride-containing media would therefore have a benign effect as found by other researchers. Roberts and Fennelly[23] showed that the presence of high Cl^- during the 1,1-TCA reduction reaction prevents the repassivation of the iron surface. On the other hand, in the elegant work of Li and Klabunde,[25] bimetallic systems were found to enhance the overall degradation rate.

4.4. AFM Images. AFM studies were undertaken to characterize the three-dimensional changes on the morphology of the iron surface with the chloride treatment and the reaction, to better visualize the defects present on the iron surface. The images shown here are only with the iron chips (TCE degradation results shown in Figure 5).

Figure 9 gives the AFM images for the fresh iron chip and the chloride-treated iron chip. The figure indicates that the surface features observed with the surface profilometer can be better viewed with AFM images. No

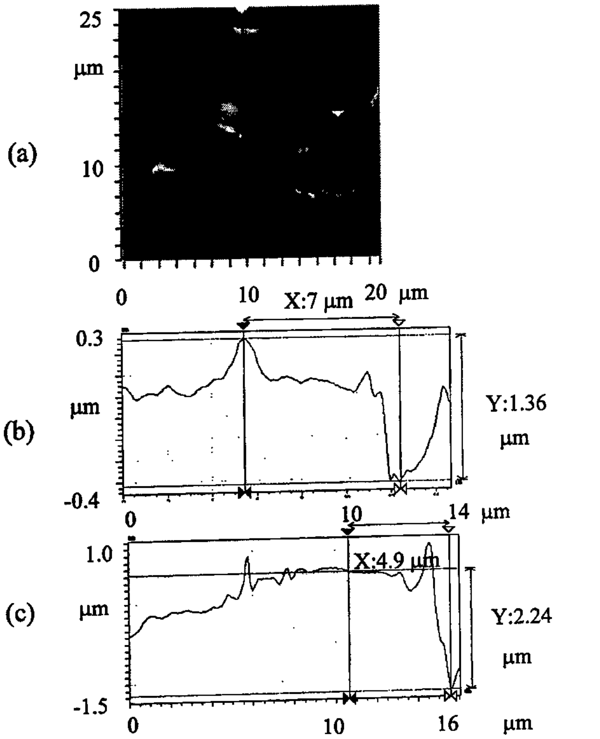


Figure 8. Surface profilometric data on the iron surface after 60 h of reaction: (a) surface image; (b) vertical profile of decreased surface roughness; (c) horizontal profile of decreased surface roughness.

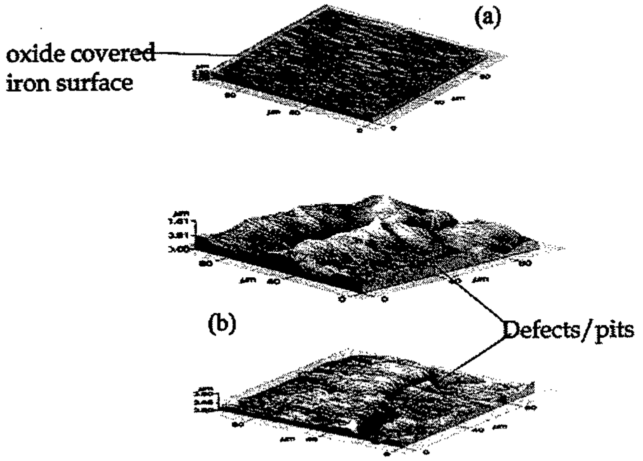


Figure 9. AFM images: (a) oxide-covered iron chip; (b) chloride-treated iron chip.

special features were observed on the oxide-covered iron surface, other than occasional dark and shiny areas as mentioned earlier. This is shown in Figure 9a. Figure 9b clearly shows the evidence of the enhanced corrosion along grain boundaries of iron as a result of the chloride treatment. We further studied the defects/abnormalities present on the surface. This is shown with the help of a two-dimensional view of the chloride-treated sample (Figure 9b), in Figure 10. It was found that the typical depth of the trenches varied from 0.1 to 0.5 μm, and the width of the trenches between the grains of iron varied from 11 to 16 μm. This is presented in Figure 10 through three profiles of a trench located between two grains of iron on the surface of the treated chip.

The effect of these surface defects and pits on the iron chips on the TCE degradation profiles is shown in Figure 5. It should be noted that the iron chips used in this study were obtained from scrap metal inventory and therefore have an oxide layer. As a result, the active external surface area of the metal is very low. Nonetheless, even with such low surface areas, the enhanced dechlorination due to defects on the surface of iron is evident from Figure 5. This again confirms the hypothesis of the defects being active sites, mainly responsible for TCE degradation.

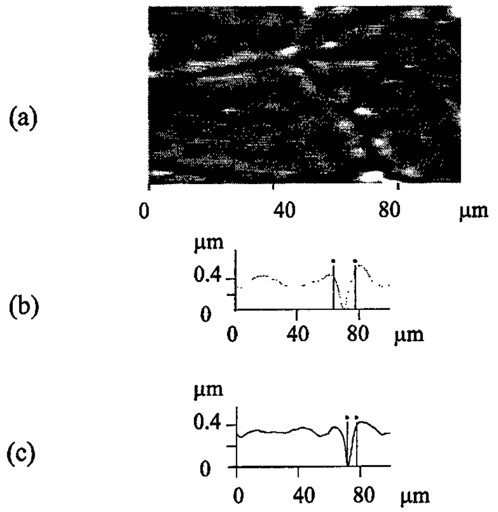


Figure 10. (a) 2-D representation of Figure 8. (b, c) Vertical profiles of trenches.

It should be noted that the surface abnormalities can be enormously increased by varying the treatment conditions.[34,35] A systematic study of these conditions will be required to achieve the optimum treatment condition, which can yield much higher degradation rates than that observed in the present study. One can also see that, as the reaction proceeds, these surface deformities diminish (Figure 8). This explains the possible reasons for the decrease in the reaction rates at longer times. However, in the presence of externally added chloride, one can enforce these deformities on the surface by continuous breakdown of the oxide precipitates that might be formed. This was confirmed by a recently published study,[23] in which much higher degradation rates of dechlorination of 1,1,1-TCA with Fe^0 in the presence of 0.1M NaCl in the reaction system were observed. Another study[33] also found that the addition of chloride ion increased the rate of CCl_4 dechlorination as high as 4-fold.

5. Possible Mechanisms for Observed Enhancement

Until now we have discussed the evidence observed for the pitting corrosion during the reductive dehalogenation of TCE with zerovalent iron. The evidence of this corrosion mechanism was further corroborated by the observed enhancements obtained by creating more pits. In this section, an attempt is made to explain the reason for this by making use of the corrosion literature.

In a recently published study, Scherer et al.,[37] have given an excellent review of the different roles that oxide layers present on the surface of iron might play during the reductive dehalogenation reaction. To explain the current results, we have used the concept of the role of the oxide layer as a physical barrier, as explained in this paper. We observed pit formation through these oxide layers when

[37] Scherer, M. M.; Balko, B. A.; Tratnyek, P. G. The Role of Oxides in Reduction Reactions at the Metal-Water Interface. In *Kinetics and Mechanisms of Reactions at the Mineral/Water Interface*; Sparks, D. L., Grundl, T., Eds.; ACS Symposium Series; American Chemical Society: Washington, DC, 1998.

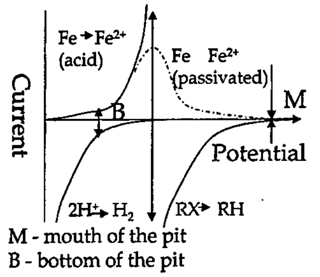


Figure 11. Possible mechanism for the observed enhancement in dechlorination rates (adapted from ref 37).

the iron was pretreated with chloride-containing solution. It should be noted that these pits create new, longer diffusion paths[35,38–40] for TCE to react with the bare metal at the bottom of the pit. Such longer diffusion paths to the bottom of the pit restrict the transport of aqueous species from the bulk of the solution. These longer diffusion paths might lower the degradation rates of TCE. However, it should be noted that two additional phenomena are encountered[37] due to pit formation. First, anodic metal dissolution in aqueous solutions containing Cl^- ion generally leads to the accumulation of metal ions and chloride ions adjacent to the metal surface. In the case of iron group metals, the accumulation of metal ions gives rise to acidification due to metal ion hydrolysis. On the other hand, relatively alkaline conditions exist at the mouth of the pit, which may favor repassivation of the surface due to Fe^{2+} diffusing from below. As with the case of crevice corrosion of iron in the presence of chloride ions, the overall rate of the process can be enhanced through the formation of surface defects, or pits.

Macroscopically, pit breakdown can be described by the current (i)−potential (E) curve. An example of a $i-E$ curve is shown in Figure 11.[37] The two half-reactions completing the electrochemical cell in both the bottom and mouth of the pit are also shown in Figure 11. As mentioned before, the conditions at the bottom of the pit are more acidic, and the half-reaction of dissolution of iron is primarily balanced by the reduction of water (because of restriction of TCE to approach the bottom of the pit due to induced longer paths). At the mouth of the pit, the half-reactions are dissolution at a passivated iron metal surface (alkaline conditions) and possible reduction of water or stronger oxidants such as oxygen or TCE. Thus, the conditions in the pit create two electrochemical cells, one at the bottom and other at the mouth of the pit. Whenever two electrochemical or galvanic cells are in close proximity to each other, the net corrosion rate of the metal is more than that with just one cell.[34] The enhancement in the degradation rate at early times could thus be attributed to the acceleration of corrosion associated with pitting results. This occurs because the close proximity of the two cells creates a coupled cell where acidic iron dissolution at the bottom of the pit and reduction of TCE at the mouth of the pit become the controlling anodic and cathodic processes. According to Scherer et al.,[37] the net rate of dehalogenation in this case would be equal to the rate of corrosion. Furthermore, the typical size of the pit observed was 11−16 μm. One of the factors governing stability of pits, or in other words preventing repassivation of pits, is the pit size. Sato[38] reports that the critical pitting radius required for stable pitting is 10−20 μm. The value we observed was also in this range, further supporting our claim about the stability of the pits. The imperfections of this order could act as initial sites for pitting.[35] The results obtained so far indicate that such a pitting mechanism seems to enhance the rate of dechlorination of TCE. In such a scenario, the major half-reaction at the mouth of the pit (in the absence of the oxygen) would be the dehalogenation of TCE. This possibility has also been speculated in a recent paper.[37]

There is, however, one more possibility for the observed enhancement. Pitting dissolution, as discussed earlier, leads to evolution of the hydrogen gas at the bottom of the pit. The possibility of this hydrogen acting as a direct reducing agent has already been discussed by Matheson and Tratnyek.[3] However, to have such action, catalysts are required. Matheson and Tratnyek[3] point out the possibility of defects/pits present on the surface acting as catalysts for such direct hydrogenation. In this study, as evidenced by AFM images and surface profilometric data, defect and pit formation was increased by chloride treatment. These defects could thus act as catalysts for the direct reduction by hydrogen generated as a result of corrosion, leading to enhancement in the reaction rate.

There is a need to do more research in this area to discern the exact mechanism responsible for the observed enhancement.

6. Conclusions

Surface characterization techniques have been employed to gain insight into the metallic surface effects involved in the reductive dehalogenation of TCE with zerovalent iron. It has been found that the defects present on the surface act as reactive sites for the dehalogenation process. A simple way to increase the number of abnormalities on the surface is by chloride pretreatment, and thus causes improvement in the degradation rates at early times. But these enhancements disappear at longer reaction times, which is attributed to the decrease in the surface roughness over the course of reaction. The increased reaction rates were attributed to the morphological changes occurring on the surface of iron, which were studied using surface profilometry and AFM. Two possible mechanisms, the proximity of two electrochemical cells and the direct hydrogenation in the presence of defects acting as catalysts have been proposed. This research explores a new approach for surface modifications that can enhance the degradation rates. Moreover, such surface modifications can also lead to reduced remediation times in in situ applications. Further analysis of the correlation between the surface defects and the degradation rates with the goal of determining optimum conditions needs to be investigated.

Acknowledgment. This study was partially supported by a Memorandum of Agreement between the Kentucky Natural Resources and Environmental Protection Cabinet and the University of Kentucky through the Kentucky Water Resources Research Institute. S.L. and R.C. acknowledge the support of NASA cooperative agreement NCC5-222, ONR Grant N00014-96-1296, and NSF Grant ECS-9724371.

Nomenclature

A_S	fractional active site concentration
b	Langmuir adsorption isotherm parameter
C^S_{TCE}	total concentration of TCE on the iron surface [nmol/g]

[38] Sato, N. *J. Electrochem. Soc.* **1982**, *129*, 260.

[39] Hassan, S. M.; Wolfe, N. L.; Cipollone, M. G.; Burris, D. R. *Prepr. Extended Abstr. Am. Chem. Soc.* **1993**, *33*.

[40] Roberts, A. L.; Totten, L. A.; Arnold, W. A.; Burris, D. R.; Campbell, T. J. *Environ. Sci. Technol.* **1996**, *30*, 2654.

C^{S*}_{TCE} concentration of TCE on the iron surface at the active sites [nmol/g]

C^{W}_{TCE} concentration of TCE in the aqueous phase [nmol/ml]

k Lagmuir adsorption isotherm parameter

$k_{intrinsic}$ intrinsic value of the degradation constant of TCE with zerovalent iron [h^{-1}]

k_{obs} observed value of the degradation constant [h^{-1}]

M Langmuir adsorption isotherm parameter

m_{Fe} amount of iron used in the reaction system [g]

V_W volume of the aqueous phase [L]

LA990325X

OPTICAL REVIEW Vol. 6, No. 4 (1999) 345–354

Nanolithography Considerations for Multi-Passband Grating Filters

Robert W. COHN,[1] Sergei F. LYUKSYUTOV,[1] Kevin M. WALSH[2] and Mark M. CRAIN[2]

[1]*The ElectroOptics Research Institute,* [2]*Electrical Engineering, University of Louisville, Louisville, Kentucky 40292, USA*

(Received November 5, 1998; Accepted February 19, 1999)

The placement accuracy and resolution of direct-write patterning tools, in particular the atomic force microscope (AFM), is considered for application to fabricating multi-passband integrated optical filters. Because of its simpler fabrication a grating structure is proposed that consists of identical stripes that are non-periodically spaced. The recently developed pseudorandom encoding method from the field of computer generated holography is modified to effectively assign analog reflectances at each point along the grating by selective withdrawal and offsetting of the stripes from a periodic spacing. An example filter designed by this method has two 1.5 nm bandwidth passbands and −23 dB of rejection for lightly coupled stripes. As with single band filters, the passbands broaden as the coupling increases. A calculation of the coupling coefficient of stripes on a fundamental mode, slab waveguide indicate that stripes on the order of 100 nm in depth and width support low insertion loss, multipassband filtering applications at visible wavelengths. Lines of these dimensions patterned with an AFM on (110) silicon indicates the feasibility of fabricating these filters. These conclusions are specific to current AFM's that are limited to writing fields of 100 μm. Increased rejection and decreased passband widths will result from incorporating precise field-stitching into future AFM's.

Key words: atomic force microscopy, nanolithography, photonic crystals, optical information processing, waveguide optics, nanometer optics

1. Introduction

Periodically spaced arrays (Fig. 1(a)) are known to strongly reflect plane waves of specific temporal frequencies determined by phase matching between the wave vector and the grating period. These structures have been applied as filters in distributed feedback laser diodes, distributed Bragg reflector fiber optic filters, planar integrated optics and volume holography. In the earliest implementations of these devices, it was common to interfere two plane waves in photosensitive films such as photoresists, photographic film, or photorefractive media to produce gratings having single wavelength reflection passbands. However, a much more general range of frequency responses is available by individually setting the position and reflectivity of each reflector in a grating (Fig. 1(b)). For example, filters that have multiple passbands can be designed, and it even is possible to specify different levels of attenuation and bandwidth for each passband. The generalized filter functions provide important building blocks for wavelength multiplexing, demultiplexing, sorting and routing functions for fiber communications systems.

Since such reflector spacings are non-periodic, the original interferometric exposure methods cannot be used and more general patterning methods are required. These patterning methods require placement precision and feature sizes that are finer than for periodic gratings. For gratings designed for visible laser wavelengths, the pitch of a periodic grating can be on the order of 250 nm (which corresponds to half the wavelength at the center frequency of the grating). Therefore resolution and line widths several times finer than the optical wavelength of interest are required to fabricate generalized non-periodic gratings.

Patterning systems with precision 2 to 3 orders of magnitude finer than visible wavelengths already exist and can provide essentially arbitrary control over the grating structures. For example, today's highest performance electron beam pattern generators direct-write lines as small as 30 nm.[1)] However, commercially available surface profiling microscopes (SPM's) also have placement resolution finer than 1 nm within a field of view of 100 × 100 μm. Various proximal probe writing methods have been demonstrated using surface profiling microscopes (including atomic force, surface tunneling and near-field optical scanning microscopes) and line widths as small as 10 nm have been reported.[2)] Furthermore, the increasing availability and the lower cost of SPM's make it is reasonable to consider their application for direct-write nanometer-scale lithography—especially during the development and prototyping of devices where writing speed of the SPM is not a critical concern.

While SPM's can provide nearly complete analog control of grating parameters, it is usually desirable if the number of fabrication variables can be reduced. This can accelerate the development, verification and, especially, the calibration of the fabrication processes. Achieving this partial control then establishes the level needed to begin developing more extensive analog control of the device parameters. Following this basic philosophy we introduce a simplified device (Fig. 1(c)) that differs from a periodic grating in the following ways: (1) Rather than locating the reflective stripes on half wavelength spacings, the stripes are placed on quarter wavelength spacings; (2) Rather than placing a stripe at each half

E-mail: rwcohn01@ulkyvm.louisville.edu

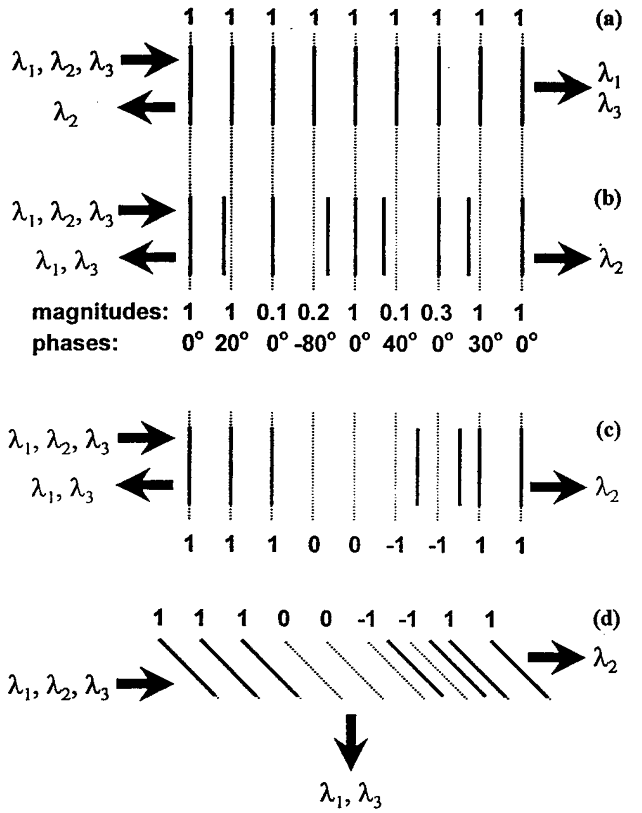


Fig. 1. Types of reflection filters. (a) Periodic reflectors which cause unit amplitude reflections that are separated by optical path differences of period λ_0, (b) aperiodic reflectors that through offsets and variable reflection strengths represent arbitrary complex valued reflectances, (c) proposed aperiodic structure that through $\lambda_0/2$ offsets and pseudorandom encoding algorithms represents the continuum of real values between -1 to 1, (d) proposed aperiodic structure with tilted reflectors. The dotted lines indicate the sampling grid for the periodic filter.

wavelength position, there is a mathematical prescription for writing, or not writing a stripe on each quarter wavelength spacing. The second difference provides a mechanism for effectively realizing a desired analog valued reflectance without resorting to varying the width or depth of individual stripes. Therefore, the new grating structure is also simplified over a fully analog aperiodic grating in that (1) the stripes are located on quarter wavelength centers rather than positioned anywhere on a continuum and (2) the reflectivity of all the stripes are identical rather than varied in an analog fashion. These fabrication constraints provide adequate flexibility to demonstrate multipassband filters.

Therefore, the main objective of this paper is to show a fabrication efficient method of designing multipassband grating filters. A second objective is to show how the design of grating filters specific to fabrication constraints can be further generalized. This philosophy and approach to filter design is analogous to the methods from the field of computer generated holography.[3-5] Specifically, the temporal frequency response of a grating is mathematically similar to the spatial frequency response (i.e. the far-field diffraction pattern) of a monochromatically illuminated computer generated hologram (CGH). This similarity can be used to directly encode a desired complex-valued temporal function/impulse response into a grating. The Fourier transform of this function is the desired temporal frequency response. These similarities are used in Sect. 2 to adapt CGH methods, specifically the recently developed pseudorandom encoding methods,[6-9] to the design of multipassband filters. In Sect. 3 a specific dual passband filter is specified using the CGH algorithm and the range of validity of the design is evaluated using a coupling of modes analysis that models the frequency response of the grating when it is implemented as a slab waveguide. The analysis accounts for multiple reflections in the grating as a function of width and depth of the grating stripes. Sect. 4 describes our initial efforts at fabricating a nonperiodic grating using an atomic force microscope (AFM). This grating provides a physical example of the proposed grating structure.

2. CGH Algorithm for Multipassband Grating Filters

From the beginnings of computer generated holography[3] until today[4,5] a critical issue has been how to represent complex valued spatial modulation with devices that do not produce arbitrary complex valued modulation. In this field the cost of implementing fully complex spatial light modulators has been considered to be difficult and costly. For this reason numerous methods of encoding fully complex valued modulation have been explored and developed specific to the modulation properties of various media. Some general classes of modulating devices include amplitude-only, phase-only, and various degrees of coupling between amplitude and phase.[6] Another classification is if the modulation values at an individual point are continuous or discrete.[7] These and other factors, as well, have stimulated many novel methods of encoding complex valued functions. The CGH design problem, in its similarity to the grating filter design problem, offers a useful source of ideas and insight for developing encoding schemes suited to the fabrication constraints of grating filters. This section adapts the CGH methodology to the problem of representing arbitrary complex-valued reflectances with the minimum increase in fabrication complexity (e.g. positioning accuracy and stripe resolution) over that needed to produce periodic grating filters.

2.1 *Complex-Valued Gratings*

Consider the case, illustrated by Fig. 1(b), of a plane wave incident on an array of reflective stripes. In this section we consider the grating to be weakly reflective so that the effect of multiple reflections can be ignored. Then the frequency response of the grating's impulse response is known to be

$$F(\nu)=\sum_{i=1}^{N} a_i \exp(j2\pi\nu t_i), \qquad (1)$$

where a_i is the (real-valued) reflectance, t_i is the time delay produced by the i'th stripe of the N stripe grating and ν is the temporal frequency of the light. The stripe reflectance can be interpreted as being complex-valued by rewriting Eq. (1) using the definitions $\nu\equiv\nu_0+\delta\nu$ and $t_i\equiv it_0+\delta t_i$ where δt_i are the offsets of the stripes it_0 from a perfectly periodic grating, and $\delta\nu$ is the frequency offset from the center frequency ν_0. Multiplying out these terms in the argument of the exponential in Eq. (1) yields a product of four complex exponentials. One term is $\exp(j2\pi\delta\nu\delta t_i)\approx 1$ for frequency ranges of concern $\delta\nu\ll\nu_0$. This condition is usually easy to meet in current wavelength division multiplexing systems where laser tuning ranges and system bandwidths are usually less than 100 nm. Ignoring this term leads to Eq. (1) being approximated as

$$\boldsymbol{F}(\nu)\approx\sum_{i=1}^{N}a_i\exp(j\varphi_i)\exp(j2\pi it_0\nu), \qquad (2)$$

where $\varphi_i\equiv 2\pi\nu_0\delta t_i$ is the nominal phase shift produced by offsetting the stripe positions from those of a periodic grating. Equation (2) is the Fourier transform of a periodic grating in which the stripes have fully complex valued reflectances. It is mathematically identical to the far-field diffraction pattern of the original Lohmann CGH,[3] in which case ν would represent the spatial coordinate across the diffraction plane. For the wide range of optical frequencies over which Eq. (2) is valid, nearly arbitrary frequency responses can be designed based on the values selected for the number of stripes, and the magnitude and phase of the stripe reflectances.

2.2 Selection of Stripe Reflectances for Dual Passband Filters

Following the CGH design philosophy, the first step in a design is to identify the available modulation values that can be implemented. Then an encoding scheme is developed to represent all the modulation values needed to design a spectrum. In this section we apply this approach to the design of a dual-passband grating filter. As discussed in Sect. 1, it is desirable to achieve the filter function with the simplest fabrication processes possible. For this reason we have specified a lithography in which each stripe is identical in geometry and stripes are written on a periodic grid corresponding to a sample spacing of $\lambda_0/2$ optical path difference where λ_0 is the wavelength at center frequency ν_0. This prescription allows reflectance values of 1, 0 and -1 to be implemented. From these values an encoding method is developed that effectively realizes a continuum of reflectances from -1 to 1. This particular algorithm is by no means the only possible CGH algorithm that could be employed, but its numerically simple implementation makes it especially useful for purposes of illustration.

In passing we note that the particular encoding method can be generalized from real, to complex-valued representations if the pitch of the sampling grid on which the reflectors are placed is reduced from $\lambda_0/2$ to $\lambda_0/3$. Then the CGH method of ternary-valued encoding can be applied with consequent improvements in the accuracy of the encoding to approximate the desired spectrum.[7] Even finer placement resolutions lead to even more accurate encoding methods. The writing of thinner lines is also desirable in that the SPM writing speed can be increased. However, thin lines must be etched more deeply to produce reflection strengths equal to those of thicker lines. The relationship between etch depth and reflectance is considered further in Sect. 3. For the frequency responses developed here, positive and negative real valued modulation is sufficient and provides the least strict requirements on line width and placement accuracy.

Based on the above considerations on the dual passband filter, each stripe will be limited by the fabrication process to be identical. Therefore a_i the magnitude of the reflectance of each stripe is identical. However, there is the option to not place a stripe at certain locations on the sampling grid. Therefore, either a unity amplitude "1" or a zero amplitude "0" can be realized at each sample point of the grid. In general, any phase $\varphi_i=2\pi\delta t_i/t_0$ can be realized by offsetting/delaying the stripes from the λ_0 sampling points on the grid. We however limit the offsets to 0 or $\lambda_0/2$. Therefore, the phases of the reflectances can be either 0 (for $\delta t_i=0$) or π (for $\delta t_i=t_0/2$). The complex reflectances $\boldsymbol{a}_i\equiv a_i\exp(j\varphi_i)$ in Eq. (2) that can be realized are "1", when a stripe is written at the $i\lambda_0$ sampling point of the grid, "-1" when a stripe is written at the $(i+(1/2))\lambda_0$ sampling point on the grid, and "0" when a stripe is written at neither of the two sampling points. [Stripes also could be written simultaneously at both the $i\lambda_0$ and $(i+(1/2))\lambda_0$ locations, but we do not consider this possibility here.]

2.3 Pseudorandom Encoding: A CGH Algorithm for Encoding Fully Complex Values

One recently developed class of CGH methods that can be adapted to the problem of encoding continuous valued reflectances with only the three amplitudes -1, 0 and 1 is referred to as pseudorandom encoding.[6,8] A specific algorithm already developed for the case of bi-magnitude SLM's will be used.[9] Given two available values of magnitude 0 and 1, bi-magnitude pseudorandom encoding can represent/encode any desired magnitude a_{ci} between 0 and 1. Used together with the additional sign reversal available by offsetting a stripe a half wavelength, all desired amplitudes between -1 and 1 can be encoded. The basic algorithm and the results of a theoretical performance analysis are given here. References 6–9 may be consulted for additional background and theory on pseudorandom CGH algorithms.

In pseudorandom encoding the magnitude a_i for the i'th stripe is selected using a random number generator. Specifically, the random number generator is configured to produce random numbers from the probability density function (pdf)

$$p(a_i)=a_{ci}\cdot\delta(a_i-1)+(1-a_{ci})\cdot\delta(a_i), \qquad (3)$$

where $\delta(\cdot)$ is the Dirac delta function, and a_{ci} is the probability of selecting the magnitude to be $a_i=1$ and $1-a_{ci}$ is the probability of selecting the magnitude to be $a_i=0$. The expected value of the random variable that has the pdf in Eq. (3) is

$$\langle a_i \rangle = 1 \cdot a_{ci} + 0 \cdot (1-a_{ci}) = a_{ci}, \tag{4}$$

where $\langle \cdot \rangle$ is the expectation operator. This shows that for bi-magnitude random selection that the probability of selecting a 1 is identical to the desired magnitude a_{ci}. Therefore, any value of a_{ci} between 0 and 1 can be realized by using a binary random number generator to select a 1 stripe with a relative frequency a_{ci} and a 0 stripe with a relative frequency of $1-a_{ci}$. Evaluating the Fourier transform of the expected grating reflectance[6–9] shows that the on-average frequency response is

$$\langle F(\nu) \rangle = \sum_{i=1}^{N} \boldsymbol{a}_{ci} \exp(j2\pi i t_0 \nu), \tag{5}$$

which with $\boldsymbol{a}_{ci} \equiv a_i \exp(j\varphi_i)$ is identical in form to Eq. (2). Thus in an average sense, pseudorandom encoding produces desired frequency responses. The quality of the encoding method is understood by evaluating the expected power spectrum, which is found to be

$$\langle \boldsymbol{I}(\nu) \rangle = |\langle \boldsymbol{F}(\nu) \rangle|^2 + \sum_{i=1}^{N} [a_{ci}(1-a_{ci})]. \tag{6}$$

The second term of Eq. (6) indicates that each stripe contributes an identifiable amount of noise. The most noise is 0.25 (when $a_{ci}=0.5$). The noise contributions approach zero as the values of the desired magnitudes a_{ci} approach either 1 or 0.

It should be noted that random bi-magnitude selection was applied previously to surface acoustic wave filters. Specifically, this invention is referred to as the withdrawal weighted interdigital transducer.[10] However, the principle of pseudorandom encoding is much more general and can be applied to a near infinite variety of modulator characteristics as is illustrated in Refs. 6–9.

3. Design and Coupled Mode Analysis of a Multipassband Filter

In this section the bi-magnitude pseudorandom encoding algorithm is applied to the design of a dual passband filter. Then the design is validated by evaluating it with a coupling of modes analysis that incorporates the effects of multiple reflections. Finally the influence of stripe width and depth on filter insertion loss is evaluated.

3.1 Fabrication Constraints on the Design

The impulse response of the grating is designed to cover optical path differences of $512\lambda_0$ or an $\sim 100\,\mu$m field of view for $\lambda_0=640$ nm. For an effective refractive index of $n_e=1.5$ a grating filter of the form of Fig. 1(c) would occupy 109 μm. Thus λ_0 optical path difference would correspond to a pitch of $\Lambda=213$ nm. However, since the layout in Fig. 1(c) permits stripes (say a -1 and a 1 in sequence) to be written as close together as $\Lambda/2=107$ nm. Therefore, considering limits on making perfectly vertical sidewalls, line widths of even less than 100 nm are generally required for the bi-phase grating.

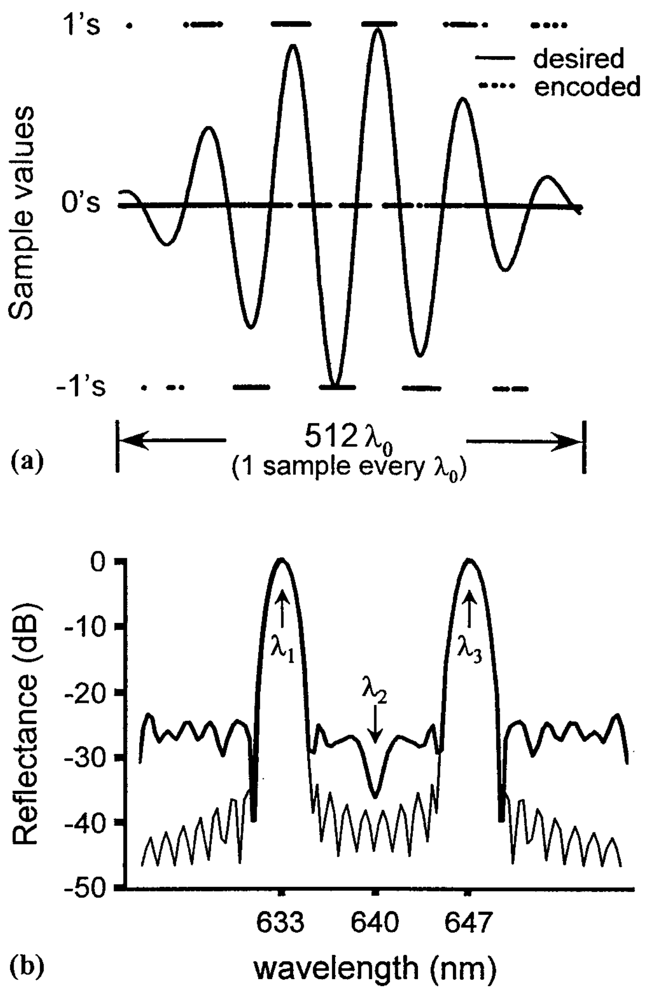


Fig. 2. Pseudorandom encoding of dual passband filter. (a) Desired real-valued reflectances and the desired function encoded with the three available values -1, 0 and 1. (b) The reflectance power spectra derived from the Fourier transforms of the temporal functions in (a). The thin line is the spectrum for the desired function and the thick line is the spectrum for the encoded function.

3.2 Grating Specification and Encoding

Based on the AFM field of view constraint we choose to design a dual passband filter that consists of $N=512$ reflectances $\boldsymbol{a}_i$. The reflectances are proportional to 1, 0 or -1 where the negative value is produced by using a $\lambda_0/2$ retardation to introduce a π phase reversal. Continuous real valued magnitudes a_{ci} are encoded using the pseudorandom algorithm from Sect. 2.

The function that is encoded is the continuous curve in Fig. 2(a). This function is a modified Dolph apodization multiplied by a sinusoid. The Fourier transformed spectrum of this function is shown in Fig. 2(b). The sinusoidal modulation introduces two passbands centered ±7 nm around the center frequency 640 nm. The Dolph apodization is known to reduce the sidelobes the greatest amount for a given broadening of the passband.[11] The Dolph function is infinite in extent, but here it has been truncated at a maximum magnitude of 0.077. This sacrifices the sidelobe level somewhat, but for a fixed field of view grating the passband is narrower than if the Dolph weights were allowed to decay to near zero. The modified Dolph apodization produces a -35 dB sidelobe level and a -3 dB bandwidth (i.e. full width at half maximum power) of 1.5 nm. This can be compared with the frequency response for 512 periodically spaced, unit strength reflectors. The periodic filter would have a sidelobe level of -13 dB and a passband bandwidth of approximately $\lambda_0/N=1.25$ nm, however the actual -3 dB bandwidth found numerically is 1.1 nm.

The encoded function is represented by the dots of values 1, 0 and -1 in Fig. 2(a). As prescribed by pseudorandom encoding, desired values close to 1 are usually, but not always, represented by $a_i=1$. Likewise values close to 0 and -1 are most frequently represented by those values. The Fourier transform of the encoded values $\boldsymbol{a}_i$ produces the spectrum (thick line) in Fig. 2(b). The bandwidth of each passband is 1.5 nm and the highest sidelobe level is -23 dB. The sidelobe level reflects the noise introduced by the noise term (i.e. the summation) in Eq. (6). The average noise level calculated from this term is -24 dB below the peak of the passband. Thus the sidelobes from the apodization are low enough that the noise from the encoding procedure is the principal contributor to the sidelobe level. The sidelobe level can be improved by using more samples in the filter. This could be achieved by designing for even shorter wavelengths or by increasing the field over which the patterning tool can write. Increasing the writing field would also allow the passbands to be narrowed further.

3.3 *Evaluation of the Grating Filter in a Slab Waveguide Configuration*

The frequency response of a periodic corrugated waveguide has been analyzed using coupling of modes (COM) analysis by Kogelnik.[12] Closed form COM solutions for non-periodic linear and quadratically chirped gratings were also developed by Kogelnik.[13] However, a method of analyzing general nonperiodic structures is needed. The analysis of arbitrary nonperiodic gratings should be analogous to the Born and Wolf analysis of a stack of nonidentical etalons.[14] Kogelnik has already adapted their method to propagation in a layered or stratified waveguide.[12] Instead of propagation being parallel to the layers we consider the case of plane wave propagation normal to the layers. We only discuss the TE case, in which case the electric field is parallel to the stripes.

While we use the analysis to evaluate the spectrum of the dual passband filter, it does not by itself provide information on the dependence of grating reflectance on the stripe width and depth. However, it is possible to relate the coupling coefficient κ used in COM analysis of sinusoidally perturbed guides to the refractive index difference Δn between the two types of layers used in a periodic etalon stack. Furthermore κ for a square wave grating on top of a slab waveguide has been directly related to that for a sinusoidal grating.[15] These relationships are used to estimate the appropriate stripe geometry as a function of the magnitude of the grating reflectance. The remainder of this section summarizes these analysis procedures and uses them to evaluate the performance of the dual passband filter design.

3.4 *Discrete Layer Analysis*

The wavelength dependent reflectance of an etalon stack can be analyzed by cascading the reflection and transmission properties of the individual layers. For this Discrete Layer (DL) analysis, each layer is modeled using a 2×2 characteristic scattering matrix as described in Ref. 14. Each section is designed to introduce a quarter wavelength optical delay $ln=\lambda_0/4$ where n is the refractive index of a particular layer and l is its physical length. The characteristic matrices for all the layers are multiplied in sequence and the resulting matrix is evaluated to give the frequency dependent complex reflectance.

For the proposed grating filter the sections that represent values of either 1, -1 or 0 are modeled as follows. Each section consists of two quarter wavelength layers. A 0 corresponds to two layers of refractive index $n=1$. A 1 corresponds to a layer with $n>1$ followed by a layer $n=1$. A -1 corresponds to a layer with $n=1$ followed by a layer with $n>1$, which is the reverse of the ordering used for the value 1. In this way the higher index layers represent stripes located at the desired positions in the proposed grating filter.

This method is adequately general for analyzing aperiodic structures. However, in order to relate the DL analysis to COM analysis it is useful to consider the special case for a periodic structure. Kogelik's analysis of the periodic, sinusoidally perturbed waveguide of length L and coupling coefficient κ gives essentially identical results as the DL analysis of N pairs of quarter-wavelength layers that differ in refractive index by Δn if

$$\Delta n\equiv\kappa\lambda_0/2. \tag{7}$$

We have also checked this correspondence through numerical simulation. We specifically evaluated a 512 period structure as a function of Δn for both analyses. The DL geometry consists of 1024 layers of alternating refractive index $n+\Delta n$ and n. The correspondence between the two models is compared in (Fig. 3) in terms of the -3 dB bandwidth as a function of filter reflectance at center wavelength $\lambda_0=640$ nm. The results are identical for the two analyses except for small errors that are due to the small number of sample points used in calculating the spectra. The bandwidth broadening is a direct result of strong multiple reflections that saturate the frequency response around the center frequency. In each case identical centerband reflectances are found if the stacked etalon and COM analyses use values of Δn and κ that are related by Eq. (7). These correspondences between the two analyses indicate that the DL analysis for the aperiodic filters will reasonably model our grating structures of interest.

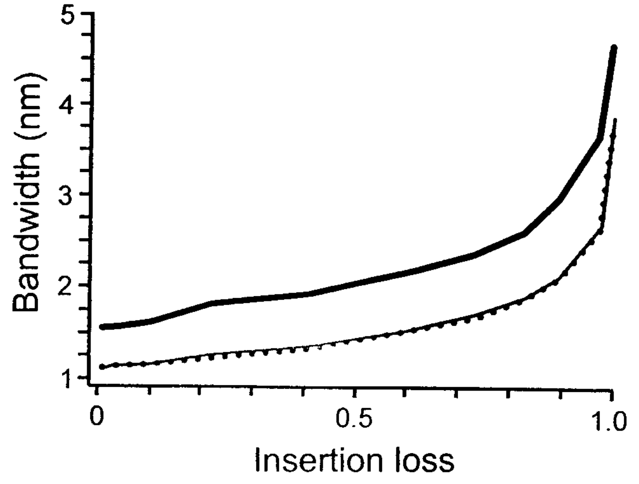


Fig. 3. Bandwidth at −3 dB of peak intensity for a single passband (512 period) grating filter and the dual passband grating filter as a function of filter insertion loss. The correspondence between the COM (coupling of modes) and DL (discrete layer) analyses for the periodic structure indicates the validity of the DL analysis for the analysis of the non-periodic dual passband filter. ——, dual band filter (DL); – – –, periodic grating (DL); ······, periodic grating (COM).

3.5 *Grating Depth Analysis*

Yariv describes a method of calculating the coupling coefficient for TE waves propagating in a slab waveguide that is perturbed by a square wave corrugation.[15] Following this procedure we derived from the unperturbed field distribution a closed form expression for the coupling coefficient. For simplicity of presentation, we only report our result for symmetric slab waveguides. We find that the coupling coefficient can be expressed as

$$\kappa=\frac{\gamma_g^2\gamma_c}{\cancel{\gamma_g^2+\gamma_c^2}}\,\frac{n_c/\lambda}{2+h\gamma_c}\,\frac{\cancel{n_g^2-n_c^2}}{\cos^2(h\gamma_g/2)}\times\left[d+\frac{\sin(h\gamma_g)-\sin(h\gamma_g-2\overset{\gamma_g}{d})}{2\gamma_g}\right], \qquad (8)$$

where h is the thickness of the guiding layer, d is the corrugation depth, n_c is the refractive index of the cladding, n_g is the refractive index of the guiding region, and $n_e=\beta\lambda/(2\pi)$ is the effective index for the wave having propagation constant β. The eigenvalues that describe the field distribution of the unperturbed field are γ_g for the guiding layer and γ_c for the cladding regions. Yariv has presented an approximate expression for κ (Eq. 13.4-17 of Ref. 15). This approximation is valid only for $h(n_g-n_c)/\lambda \gg 1$. This condition is valid only if the guide can support multiple modes.[12] However, Eq. (8) is valid even for single mode guides.

Equation (8) specifically describes κ for the fundamental waveguide mode and the fundamental Fourier series harmonic of a 50% duty cycle grating. The expression of coupling coefficient is generalized for a rectangular grating of any duty cycle according to

$$\kappa_n=\kappa\sin(\pi\Delta), \qquad (9)$$

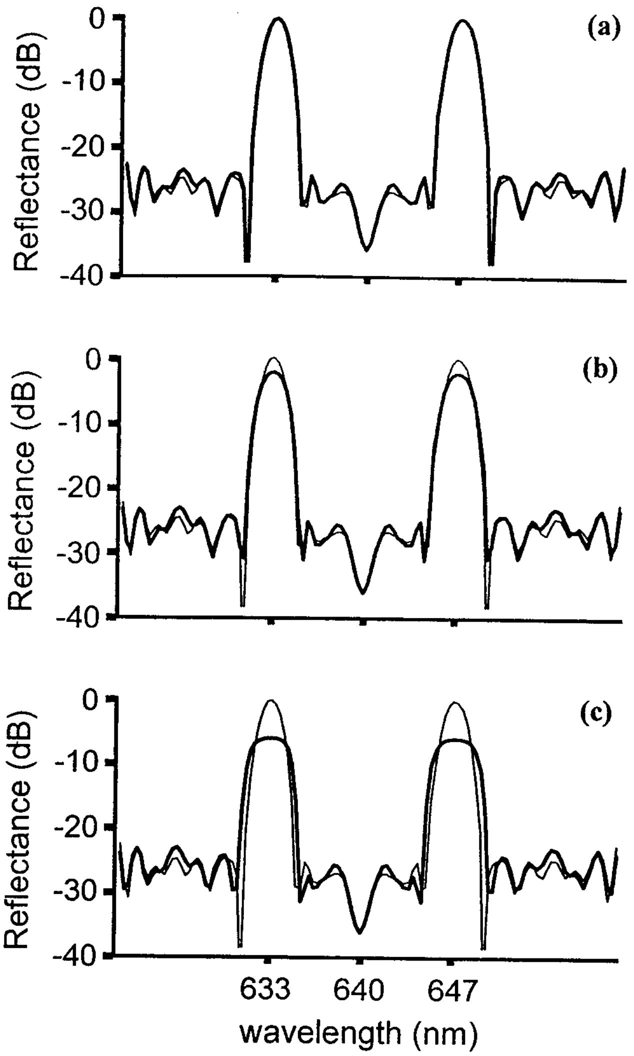


Fig. 4. Reflectance spectra for the dual passband filter design. The DL analyses (thick lines) are shown for filter insertion loss (i.e. peak intensity reflectance at centerband) of (a) 0.068, (b) 0.532 and (c) 0.917. The spectrum for the dual passband design from Fig. 2 is also replotted (thin line) for comparison. The reflectance spectra are normalized so as to bring their sidelobe structure into correspondence with the design spectrum.

where the term $\sin(\pi\Delta)$ is the ratio of the Fourier coefficient of a grating of duty cycle Δ to the coefficient for $\Delta=1/2$.

3.6 *DL Analysis of the Dual Passband Filter*

Figure 4 shows representative spectra resulting from the DL analysis for values of Δn of 0.0015, 0.0065, and 0.14 for a–c respectively in Fig. 4. For a centerband reflectance (i.e. insertion loss) of 0.068 (−11.7 dB) the DL analysis in Fig. 4(a) is nearly identical to the Fourier transform of the encoded function (from Fig. 2). For an insertion loss of 0.532 (−2.7 dB) the mainlobes in Fig. 4(b) are slightly saturated and the sidelobe are nearly identical to the designed spectrum. For insertion loss of 0.917 (−0.38) dB, Fig. 4(c) shows a strong intensity saturation and frequency broadening of the passbands.

Again, the sidelobes are nearly identical to those of the designed spectrum. These results show that encoded design methods serve a useful role even when the passbands are heavily saturated; namely, improving rejection by shaping the sidelobe region. Figure 3 summarizes the distortion from the designed spectrum in terms of bandwidth broadening for various levels of insertion loss. As with the periodic gratings the bandwidth increases with decreasing loss.

We also compared the values of Δn used in the DL analyses of the periodic grating with the values of Δn for the non-periodic grating. We found for equal insertion loss that Δn was typically 3.7X greater for the aperiodic grating than for the periodic grating. This ratio is similar to the ratio of the peak magnitude of the Fourier transform of the periodic grating to that of the non-periodic grating which is 4.0X. Furthermore, since coupling coefficient in Eq. (7) is proportional to Δn, the proportionality between Δn for the periodic and non-periodic gratings gives some idea of the stripe depth required to achieve a desired level of insertion loss.

3.7 Analysis of Stripe Geometry

Equation (8), the relationship between stripe depth d and coupling coefficient for a periodic grating of 50% duty cycle (where duty cycle is the ratio of stripe width to grating period) is evaluated in Fig. 5 for four values of guide thickness $h=0.5$, 1, 2 and 3 μm. Figure 5(a) shows the coupling coefficients for $n_g=1.5$ and Fig. 5(b) shows the coupling coefficient for $n_g=1.05$. In both cases $n_c=1$. In Fig. 5(a) the curves for $h=0.5$ and 1 μm correspond to single mode operation while all four curves in Fig. 5(b) are for single mode operation.

Figure 5(a) shows coupling coefficients as large as 50 mm^{-1}. For the 512 period periodic filter -1 dB insertion loss (0.8 reflectance) corresponds to $\kappa=8.8$ mm^{-1}. However, since the duty cycle Δ for the aperiodic grating is at most 25% to avoid overlap of stripes Eq. (9) gives that $\kappa_n/\kappa \leq \sqrt{1/2}$. Additionally, since the peak amplitude of the periodic filter is 4X less than the dual passband filter for the same value of then coupling of at least 50 mm^{-1} is required to obtain dual passband filters with -1 dB insertion loss. Figure 5(a) shows that -1 dB insertion loss is possible using stripes of depth ~50 nm for the 0.5 micron guiding layer and ~130 nm for the 1 micron layer. For the lower index guide of Fig. 5(b), a coupling coefficient of only 10 mm^{-1} is achieved for stripe depths of ~70 nm and ~125 nm. Note however that reducing κ by a factor of 14.4X reduces insertion loss from -1 dB to -20 dB. For the dual passband filter considered here, -20 dB corresponds to $\kappa=3.5$ mm^{-1}. For the Fig. 5(b) curve the stripe depths would correspond to approximately 25 nm, 50 nm and 150 nm for the 0.5, 1 and 2 μm guides respectively. The point of this analysis is that there is substantial flexibility in adjusting stripe depth and width, and guiding layer thickness to obtain low insertion loss filters, lightly coupled wavelength selective drops, and intermediately coupled power splitters.

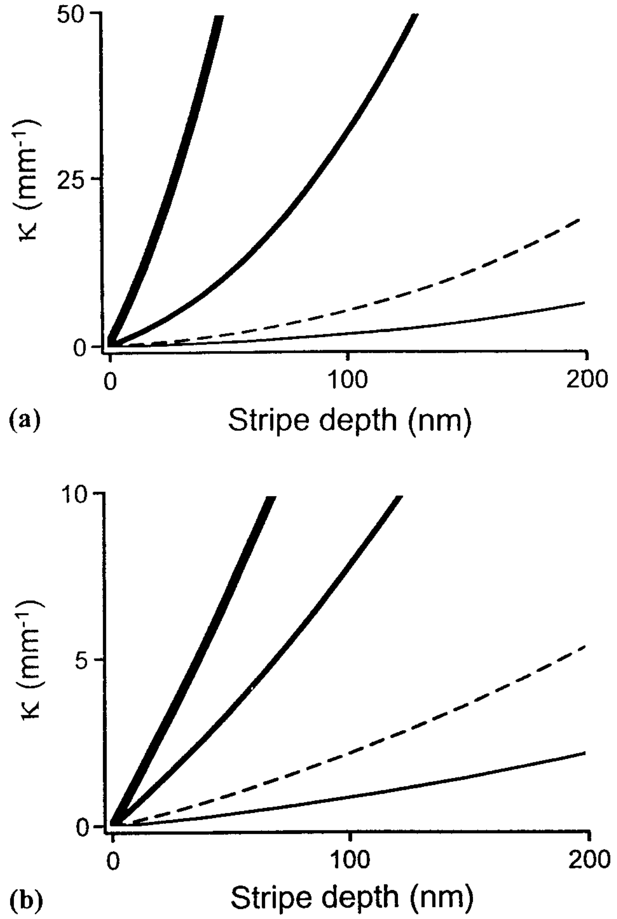


Fig. 5. Reflected wave coupling strength for a periodic grating on a symmetric slab waveguide. The results shown are for the fundamental TE mode of the guide, guiding layers of thickness between 0.5 and 3.0 μm, and a 50% duty cycle grating. The curves are for cladding index $n_c=1$ and for guide index (a) $n_g=1.5$ and (b) $n_g=1.05$. ——, 0.5 μm; ——, 1.0 μm; ----, 2.0 μm; ——, 3.0 μm.

In passing we note that somewhat narrower stripes will not require significant increases in stripe depth, though substantially narrower stripes will. For example, for a duty cycle $\Delta=1/6$ a compensation $\kappa/\kappa_n=2$ in stripe depth is needed to obtain identical reflectivity as a 50% duty cycle grating, while for $\Delta=1/32$ a compensation of $\kappa/\kappa_n=10$ is needed. These results give some idea of the tradeoff between stripe depth and stripe width. Thus, while narrower stripes are desirable in that they allow finer placement with consequent improvements in line writing speed and enhanced performance encoding algorithms, this must be traded off with the requirements for increasing depth of the stripes. These limitations can be further compensated if filters having a greater number of stripes can be fabricated. Ways that the writing range of the AFM might be extended are considered further in Sect. 5.

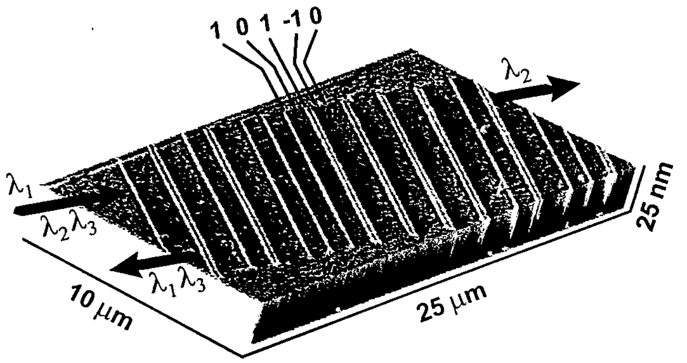


Fig. 6. A non-periodic silicon dioxide grating that has been direct written on a silicon surface using an atomic force microscope. The structure (which also was profiled by an AFM) is annotated with wavelengths and stripe reflectances in the same manner as the proposed filter of Fig. 1(c). The closest spacing of adjacent lines in this AFM profile is 240 nm for a 1 followed by a −1.

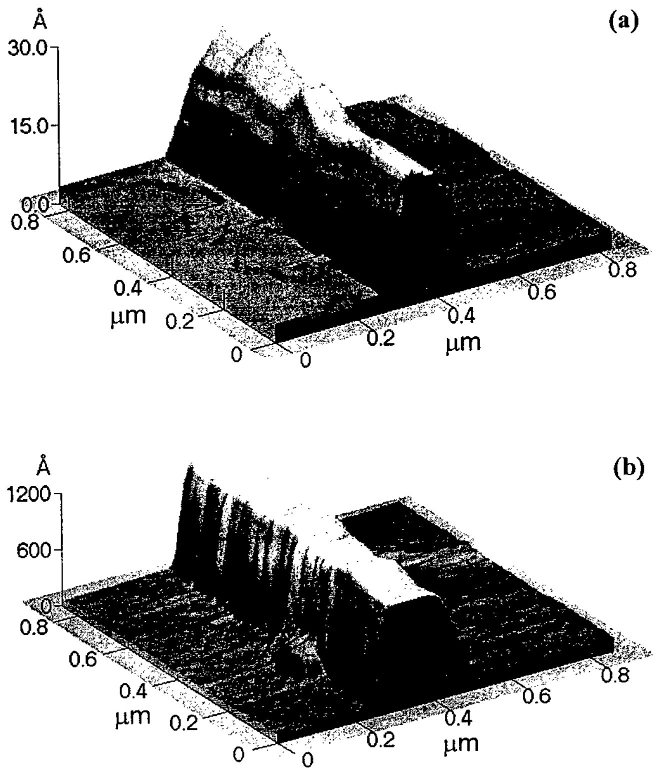


Fig. 7. Close up AFM profiles of (a) oxide line on (110) silicon and (b) same line after anisotropic wet chemical etching in TMAH.

4. Initial Fabrication of Non-Periodic Gratings for Multi-Passband Reflection Filters

Figure 6 illustrates the device concept that has been explored in this paper. The figure shows a series of non-periodically spaced silicon dioxide stripes that were written on silicon and profiled using an AFM. Stripe positions corresponding to 1, −1 and 0 filter values are indicated. The closest spacing between adjacent lines is 240 nm for a 1 followed by a −1. Figure 6 also shows how the wavelengths λ_1 and λ_3 would be separated from λ_2 for the dual passband design of Sect. 3. Of course, complete separation (or any desired division) between the transmitted and reflected channel would require that the stripes are of the appropriate height to obtain close to 100% reflectance (i.e., 0 dB insertion loss). This section will describe initial material processing experiments aimed at obtaining fabrication control over the stripe geometry.

Silicon surfaces can be oxidized by applying large electric potential to them. Various studies have shown proximal probe oxidation process using a biased surface tunneling or atomic force microscope tips.[2,16-18] Many other writing modes of surface profiling microscopes (SPM) and various material systems have been reported that could be employed for fabricating optical devices.[2] Our writing experiments are performed with (110) n-type silicon. Prior to writing the wafer is cleaned and the native oxide layer is removed by immersing the wafer in $HCl:H_2O_2:H_2O$ (3:1:1) at 70°C for 10 min followed by 20–30 s etching in a 40:1 HF solution.[19,20] The surface roughness, as measured by the AFM, is less than 0.3 nm (rms) if the wafer is processed soon after the residual oxide is removed.

The oxide lines are written with a Park M5 AFM in room air. A silicon contact mode tip (UL06) mounted on the conductive holder is biased between −5 to −10 V and the sample is grounded. The resistance between the sample surface and ground was measured to be 1000 Ω. The tip is placed in contact with the sample and then moved over the surface to write a line. The amount of oxide and the thickness of the line depends on the voltage and the amount of time that the AFM tip is in contact with the sample. After writing the oxide is profiled with the AFM.

We have written oxide lines that vary in width from 100 nm to 400 nm and that vary in height from 0.5 to 2.0 nm. The width refers to the maximum width of the base that can be detected by the AFM. This is limited by AFM height resolution to 0.08 nm. To the eye, the shape of the oxide appears to be gaussian (Fig. 7(a)). A 120 nm width oxide line was produced with an applied voltage of −10 V and a scan speed of 2 μm/s. Slower scanning speeds or multiple passes over the same region produce wider lines. These oxide lines can be used as a mask to etch substantially deeper lines.

Based on the analyses of κ in Sect. 3, it would be possible to obtain adequately strong coupling with 2 nm thick stripes over a very thin guiding layer. However, we have concerns about residual surface roughness of the wafer and for these reasons deeper stripes appear desirable. Crystalline materials, such as silicon can be anisotropically etched using wet chemical processing.[21,22] Objections to the limited number of crystalline materials available can be overcome by using three dimensional pattern transfer methods.[23] Furthermore, nearly vertical sidewalls can be produced in a much wider variety of materials by reactive ion etching.[21,22]

In our first experiments we consider anisotropic wet chemical etching of silicon. In particular we choose (110) silicon because the (110) plane is known to etch as much as 100X faster than the (111) plane for specific etchants.[21] Thus by writing oxide lines on the surface of (110) silicon that are parallel to the wafer flat we anticipate that near vertical sidewalls corresponding to the (111) plane will be formed. The AFM written silicon dioxide lines are oriented in this way.

A single wafer containing lines of various widths is wet etched at 90°C in a solution containing 83 ml of 25%wt. tetramethyl-ammonium hydroxide (TMAH) and 17 ml of isopropyl alcohol for 15 s. AFM profiles of the etched structures reveals several interesting results. (1) The thicker oxide lines etch to a depth of ~100 nm while the thinner lines etch to depths of ~50 nm. (2) The typical sidewall slope is 30° from vertical. (3) Thicker lines have a flat surface between the sidewalls while the thinner lines form a continuously curved hillock. These results suggest that the thin edges of the oxide are not protecting the line during the entire etch. For the very thin lines the oxide is being completely undercut and removed. Figure 7 shows one of the lines before (Fig. 7(a)) and after etch (Fig. 7(b)). The width across the base of the line is ~250 nm both before and after etch. After etch the width of the plateau of the line has narrowed to ~100 nm. Additional studies are needed to find ways to better protect silicon from etchants, such as modifying oxide shape or optimizing the properties of the etchant. Alternatively, different material processes altogether may lead to lines of the desired width and depth.

5. Discussion

We have considered the possibility of generalizing the frequency response of grating reflection filter by employing signal encoding techniques from computer generated holography. A specific case of a grating on a slab waveguide has been considered. Fabrication constraints set by the limited field of view of current AFM patterning systems have been considered. One consideration is that encoded functions generally have lower intensity frequency responses than do periodic structures of the same length. In order to compensate for these differences the stripes must be correspondingly deeper than for periodic structures in order to obtain equivalent insertion losses. Likewise, the use of phase reversed strip placement for the dual passband filter requires narrower stripes which also requires deeper stripes. Even greater depths (or thinner guiding layers) will be needed to extend this approach from bi-phase to polyphase encoding algorithms. Already for some of the examples considered here, stripe aspect ratios (depth over width) in excess 1:1 have been found.

These aspect ratios can be reduced for designs having more stripes. Current AFM's (unlike ebeam pattern generators) do not include high precision stages necessary to stitch together multiple fields. At least two reasonable extensions are possible:

(1) Two gratings could be placed in close proximity to each other. A short region that is unperturbed by a grating is placed between the two sections. The guided wave velocity can be compensated by depth etching, to properly phase the two sections together. A single etch depth (i.e. identical etch conditions) can be used for phasing any two sections if the length of the guided region that is exposed to etchant is varied.

(2) While commercial AFM's are quoted with around 2–3 μm of stage positioning error, it would be possible to use the AFM head itself to determine exactly where the stage moves to. The previously written pattern (or other prewritten fiducials) can be identified through AFM scanning and the measured offset (and possibly tilt) errors can be used to offset (and possibly rotate) the patterning instructions. This is a quite reasonable approach if one has the ability and adequate time to modify the AFM control software.

There are various applications and configurations of multipassband grating filters. A single grating customized to the demands of a subscriber can be used to tap off a number of non-sequential frequency channels from a wavelength multiplexed fiber channel. Slanted gratings (Fig. 1(d)) can be used to form wavelength selective crosspoints. Arranging these crosspoints an x-y fashion on a single substrate can be used to realize various other network topologies. Programmable filters can also be envisioned in which arrays of stripes can be individually placed in or removed from the beam path. Electrostatic attraction could be used to displace the stripes in a manner similar to micromechanical mirror array technology of Texas Instruments that is currently used in video projectors and printing engines.[24] CGH encoding algorithms, such as those described provide the flexibility and adaptivity to design and compute desired filter functions instead of storing large tables of anticipated stripe settings.

In summary this paper has considered the possibility of nanofabricating multipassband grating filters using current AFM's as direct write patterning tools. We have demonstrated that useful designs are possible even with the limited writing field of current AFM's and that there are approaches that can permit precise field stitching. Writing times are currently quite slow but not critical for using AFM's to develop single experimental or prototype devices. Current writing speed can be increased by using a controlled atmosphere for the silicon oxidation process or using different material systems which are known to be faster.[25]

Acknowledgments

We thank Frank Luecke of FrankDesign, LLC for discussions and ideas on field stitching methods. This study was supported by ONR grant N00014-96-1-1296, NASA cooperative agreement NCC5-222, and NSF grant ECS-9724371.

References

1) H. Asai and Y. Wada: Proc. IEEE **85** (1997) 505.
2) E. S. Snow, P. M. Campbell and F. K. Perkins: Proc. IEEE **85** (1997) 601.
3) B. R. Brown and A. W. Lohmann: Appl. Opt. **5** (1966) 967.

4) W. J. Dallas: *The Computer in Optical Research*, ed. B. R. Frieden (Springer, 1980) Chap. 6, p. 332.

5) R. W. Cohn and L. G. Hassebrook: *Optical Information Processing*, eds. F. T. S. Yu and S. Jutamulia (Cambridge U. Press, 1998) Chap. 15, p. 396.

6) R. W. Cohn: J. Opt. Soc. Am. A **15** (1998) 868.

7) R. W. Cohn and M. Duelli: J. Opt. Soc. Am. A **16** (1999) 71.

8) R. W. Cohn and M. Liang: Appl. Opt. **33** (1994) 4406.

9) R. W. Cohn and W. Liu: *1996 OSA Technical Digest Series*, **5** (Boston, MA, 1996) pp. 237–240.

10) C. S. Hartmann: 1973 IEEE Proc. Ultrasonics 1973, 423.

11) F. J. Harris: Proc. IEEE **66** (1978) 51.

12) H. Kogelnik: *Guided-Wave Optoelectronics*, 2nd ed, ed. T. Tamir (Springer, 1990) p. 7.

13) H. Kogelnik: Bell Syst. Tech. J. **55** (1976) 109.

14) M. Born and E. Wolf: *Principles of Optics* (Cambridge U. Press, 1980) 6th ed.

15) A. Yariv: *Optical Electronics* (Oxford U. Press, 1997) 5th ed, Chap. 13, p. 491.

16) J. A. Dagata, J. Schneir, H. H. Harry, C. J. Eaves, M. T. Postek and J. Bennet: Appl. Phys. Lett. **56** (1990) 2001.

17) E. S. Snow, P. M. Campbell and P. J. McMarr: Appl. Phys. Lett. **63** (1993) 749.

18) E. S. Snow, W. H. Juan, S. W. Pang and P. M. Campbell: Appl. Phys. Lett. **66** (1995) 1729.

19) H. Sugimura and N. Nakagiri: Nanotechnology **6** (1995) 29.

20) H. Sugimura and N. Nakagiri: Nanotechnology **8** (1997) A15.

21) K. E. Petersen: Proc. IEEE **70** (1982) 42.

22) G. T. A. Kovacs, N. I. Maluf and K. E. Petersen: Proc. IEEE **86** (1998) 1536.

23) Y. Xia and G. M. Whitesides: Annu. Rev. Mater. Sci. **28** (1998) 153.

24) P. F. Van Kessel, L. J. Hornbeck, R. E. Meier and M. R. Douglass: Proc. IEEE **86** (1998) 1687.

25) K. Wilder, C. F. Quate, B. Singh and D. F. Kyser: *Electron, Ion and Photon Beam Technology and Nanofabrication*, paper NT7 (26–29 May 1998, Chicago, IL).

Pseudorandom encoding for real-valued ternary spatial light modulators

Markus Duelli and Robert W. Cohn

Pseudorandom encoding with quantized real modulation values encodes only continuous real-valued functions. However, an arbitrary complex value can be represented if the desired value is first mapped to the closest real value realized by use of pseudorandom encoding. Examples of encoding real- and complex-valued functions illustrate performance improvements over conventional minimum distance mapping methods in reducing peak sidelobes and in improving the uniformity of spot arrays. © 1999 Optical Society of America

OCIS codes: 230.6120, 090.1760, 030.6600.

1. Introduction

Complex-valued spatial light modulators (SLM's) greatly simplify the design of transmittance functions for multispot beam steering systems and other Fourier-transform processors. With arbitrary complex modulation, many desired patterns can be specified with standard Fourier-transform tables. However, fully complex SLM's either are not widely available or are rather involved to construct.[1] For these reasons encoding methods are often used to approximate fully complex operation.[2–4] In adaptive or rapidly updated systems encoding may be preferred to global optimization methods[5] because of its speed. Because current SLM's have relatively low numbers of pixels compared with diffractive optics and holograms, methods that use group-oriented encoding[6,7] are undesirable in that they further reduce the useful space–bandwidth product. Two general methods that avoid grouping are pseudorandom encoding[8] (PRE) and minimum distance encoding[9] (MDE). Both methods map each desired complex value to a realizable modulation value of each corresponding SLM pixel. Most recently these methods have been evaluated and compared for SLM's that produce at least three quantized phase-only values.[10]

For the case of real-valued ternary modulation (i.e., SLM's that produce the modulation values of 1, 0, and -1) it is not immediately evident that PRE can support fully complex representations. The problem is that the range of values that can be encoded is limited to real values between -1 and 1 [Fig. 1(a)]. MDE is not limited to the real axis, since MDE maps the desired complex value to the closest modulation value. As shown in Fig. 1(b) MDE divides the complex plane into distinct regions. Any complex value in a given region is mapped to the single modulation value in that region. However, we have noted for other types of SLM that the accuracy with which the diffraction patterns approximate the desired diffraction patterns can be improved on by use of other types of encoding.[10,11]

The greatest improvement observed has been by use of a hybrid encoding algorithm that blends PRE with a modified MDE algorithm.[12] In this method the desired complex value is mapped to the closest value that can be produced by PRE. The mapped value is then pseudorandom encoded to produce the modified minimum distance PRE (mMD-PRE). This type of encoding is illustrated for the ternary SLM in Fig. 1(c). The mMD-PRE is a specific variant of PRE that permits complex-valued representation, even with the real-valued ternary SLM. The mMD-PRE can be contrasted with conventional minimum distance-PRE (MD-PRE). For the ternary SLM all values would be encoded by use of MDE [as in Fig. 1(b)] except those real values between -1 and 1, which would be encoded by PRE [as in Fig. 1(a)]. For typical complex-valued functions a negligible number of values would be encoded by PRE, and therefore the algorithm can be

The authors are with the ElectroOptics Research Institute, University of Louisville, Louisville, Kentucky 40292. R. W. Cohn's e-mail address is rwcohn01@ulkyvm.louisville.edu.

Received 22 January 1999; revised manuscript received 22 March 1999.

0003-6935/99/173804-06$15.00/0

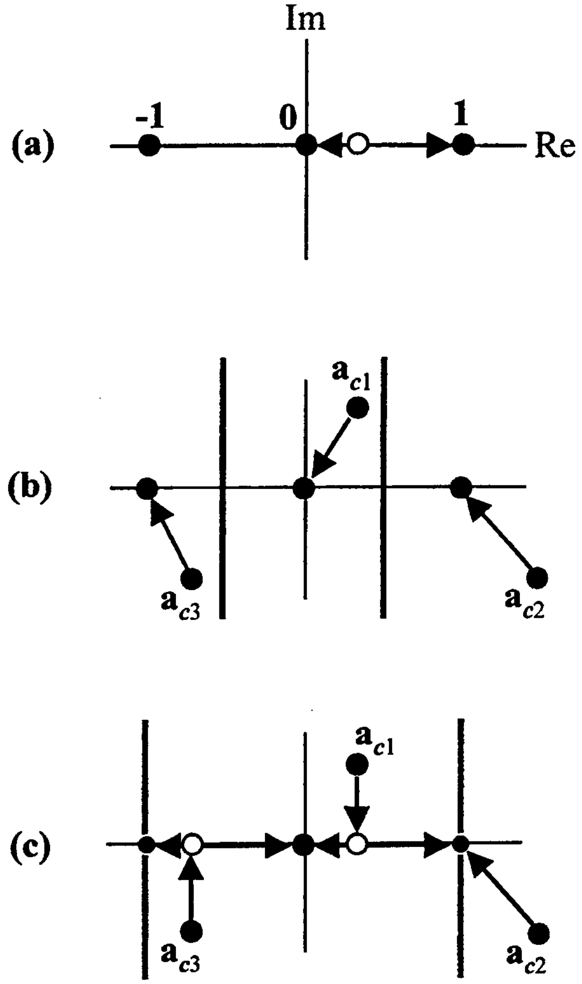


Fig. 1. Pixel-oriented encoding methods: (a) PRE, (b) MDE, (c) mMD-PRE. Method (c) also describes MD-PRE if the desired function is strictly real.

viewed essentially as being MDE. Two SLM's that produce such real ternary modulation are magneto-optic SLM's and analog chiral smectic liquid-crystal SLM's.[13,14]

Our objectives in this paper are (1) to demonstrate the feasibility of complex-valued representation by use of mMD-PRE with a real-valued ternary modulation characteristic, (2) to demonstrate the improvements over MDE alone, and (3), for the encoding of a strictly real function, to show the improvements of MD-PRE over MDE. These encoding algorithms also illustrate in brief the same general characteristics and performance trends that were reported for the application of these methods to various quantized phase SLM's in Ref. 12.

The paper is organized as follows: We present the various encoding algorithms; review definitions of the metrics used to compare them; and then compare the encoding algorithms, using computer simulations of two spot array generator designs (one in which the desired function is strictly real and the other in which the desired function is complex.)

2. Description of the Encoding Algorithms

A. Minimum Distance Encoding

The MDE algorithm maps each desired complex value $\mathbf{a}_{ci}$ (where i is the pixel index) to the closest available value (-1, 0, or 1) of the SLM. As illustrated in Fig. 1(b), the complex plane is divided into three decision regions. This mapping can be expressed as

$$\begin{aligned} \mathbf{a}_i &= \mathrm{sgn}[\mathrm{Re}(\mathbf{a}_{ci})] && \text{if } \tfrac{1}{2} \le |\mathrm{Re}(\mathbf{a}_{ci})|, \\ \mathbf{a}_i &= 0 && \text{if } |\mathrm{Re}(\mathbf{a}_{ci})| < \tfrac{1}{2}, \end{aligned} \qquad (1)$$

The performance of the resulting diffraction pattern from this transmittance function can depend greatly on the scaling of the desired function $\mathbf{a}_{ci}$. The scaling can be written as

$$\mathbf{a}_{ci} = \gamma \exp(j\beta)\mathbf{a}'_{ci}, \qquad (2)$$

where the maximum magnitude of $\mathbf{a}'_{ci}$ for $i = 1$–N is unity. It is typical to optimize the performance metric of interest as a function of the two parameters γ and β. These parameters are also used for the same purpose in the blended algorithms considered here.

B. Pseudorandom Encoding

PRE is a statistically based method of encoding in which one modulation value from a range of possible values is selected with a computer-generated random (i.e., pseudorandom) number. The statistical properties of the random-number generator are designed so that the average modulation value is identical to the desired complex value [see Eq. (1) in Ref. 10]. The diffraction pattern produced when we encode the values $\mathbf{a}_{ci}$ of the desired transmittance function has an average intensity that is identical to the desired diffraction pattern plus a noise background. Additional theory and algorithm derivation procedures for a wide variety of modulator characteristics were presented in Refs. 15–17.

C. Modified Blended Encoding

Following the two procedures above, we directly state the mMD-PRE algorithm. Any desired value found on the real axis between -1 and 1 is encoded by PRE [Fig. 1(a)]. Desired values that have real parts that lie between -1 and 1 are projected to the closest point on the real axis, and then the projected value is encoded by PRE [Fig. 1(c)]. Values that have real values that are greater than 1 or less than -1 map to the closest available modulation values 1 and -1, respectively. The mathematical specification of the encoding algorithm associates a probability with the desired value $\mathbf{a}_{ci}$ according to

$$p = |\mathrm{Re}(\mathbf{a}_{ci})|. \qquad (3)$$

With this value of probability the encoding formula is

$$\begin{aligned} \mathbf{a}_i &= \mathrm{sgn}[\mathrm{Re}(\mathbf{a}_{ci})] && \text{if } 0 \le s_i < p_i \quad \text{or } 1 < p_i, \\ \mathbf{a}_i &= 0 && \text{if } p_i \le s_i \le 1, \end{aligned} \qquad (4)$$

where $\mathbf{a}_i$ is the actual modulation selected for the ith modulator pixel and s_i is a pseudorandom number selected from the uniform distribution with mean 0.5 and a spread of unity. Equation (4) shows that the closer the real part is to an actual modulation value the more frequently that value is selected. For cases in which the real magnitudes exceed unity, p_i cannot be considered to be a probability and random selection cannot be used. Instead MDE is used that corresponds to the first line of Eq. (4) when the second part of the if statement is true. Also note that, if the desired function is strictly real, then Eqs. (3) and (4) also describe MD-PRE. That is, the real values between -1 and 1 are encoded by PRE, and the values with magnitudes greater than unity are encoded by MDE. In this study both strictly real and fully complex desired functions were encoded, which permits comparisons of mMD-PRE with MDE and PRE individually and comparisons of MD-PRE with MDE and PRE individually.

3. Design of Simulation Experiments

The real-valued and the fully complex desired functions are designed to produce a 7 × 7 array of uniform intensity spots in the diffraction plane. They are based on the functions reported in Tables 1 and 3 of Krackhardt *et al.* for 1 × 7 spot arrays.[18] Their functions on conversion to biamplitude (1, -1) and analog phase-only functions have the highest possible diffraction efficiency (their transform from desired function to realizable modulation is equivalent to MDE with $\gamma = \infty$). Our two desired functions are two-dimensional rectangularly separable functions that are constructed when we cross their one-dimensional functions. The function is sampled to produce a 128 × 128 pixel matrix that consists of a 4 × 4 array of unit cells. Additionally, a phase ramp is added to the fully complex function so that it reconstructs off axis. The phase ramp makes the encoded function essentially independent of the phase parameter β. Therefore only γ is varied in the evaluation of the encoding algorithms. For purposes of evaluating the encoding algorithms with real-valued functions, again, only γ is varied.

The key metrics of interest describe the accuracy (or fidelity) with which the actual reconstruction matches that for the desired function. These are the nonuniformity (NU) of the spot array, which is calculated as the standard deviation of the peak intensities of the 49 spots divided by the average intensity of the spots, and the signal-to-peak-noise ratio (SPR), which is the ratio of the average peak intensity of the spots to the maximum noise peak found in the diffraction pattern (excluding the square region that contains the 7 × 7 spot array). We feel that these metrics are especially important in real-time systems for which it is not practical to perform designs on the fly with numerically intensive optimization.

It is also common to report diffraction efficiency η for most designs. We can calculate this by first summing the intensities of the 49 spots, dividing by the sum of all intensities in the diffraction pattern, and then multiplying this by an additional factor that accounts for the absorption by the zero-valued modulation states. This factor is simply the ratio of the unity magnitude values divided by the total number of SLM pixels. We will show that there is a continuous trade-off between fidelity and diffraction efficiency and the best fidelity is achieved when the diffraction efficiency is less than the maximum possible.

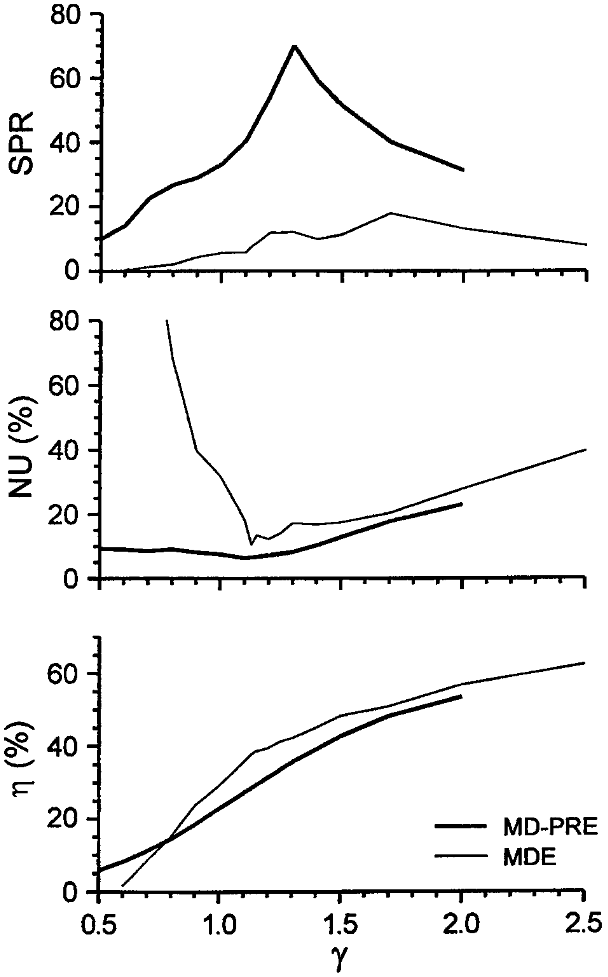


Fig. 2. Performance of MDE and MD-PRE as a function of the magnitude scaling parameter for encoding the real-valued function.

4. Comparison of the Encoding Methods

Figure 2 shows how the performance of encoding the

Table 1. Performance for Encoding the Real-Valued Function

Encoding	γ	η (%)	SPR	NU (%)
MD-PRE[a]	1.1	27	41	6.3
MDE[a]	1.13	38	7.9	10.4
MD-PRE[b]	1.3	36	70	8.1
MDE[b]	1.7	51	18	20.3
MDE	∞	73	5.1	45.0

[a] Minimum NU.
[b] Maximum SPR.

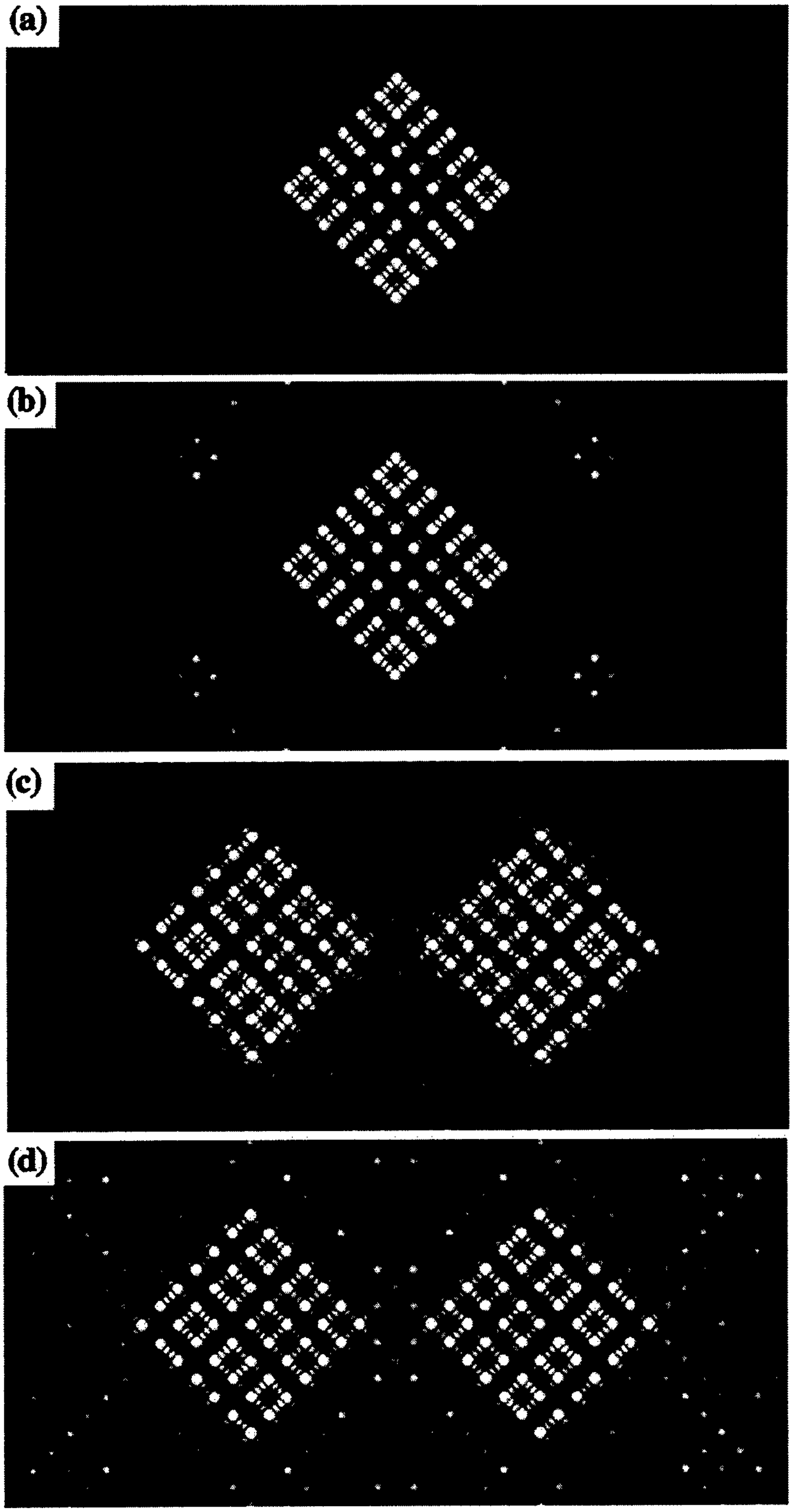


Fig. 3. Diffraction patterns for (a) MD-PRE for $\gamma = 1.1$ and (b) MDE for $\gamma = 1.13$ for the real-valued desired function and (c) mMD-PRE for $\gamma = 1.05$ and (d) MDE for $\gamma = 1.9$ for the complex-valued desired function. The intensity images are saturated so that the full white gray scale corresponds to 1⁄10 of the average intensity of the 49 spots. Also, the images are shown rotated by 45° from the x–y coordinate system.

real-valued function by use of MDE and MD-PRE depends on the parameter γ. For both encoding algorithms the largest value of SPR and the smallest value of NU are found for γ somewhat greater than unity. MD-PRE always achieves significantly larger values of SPR and somewhat smaller values of

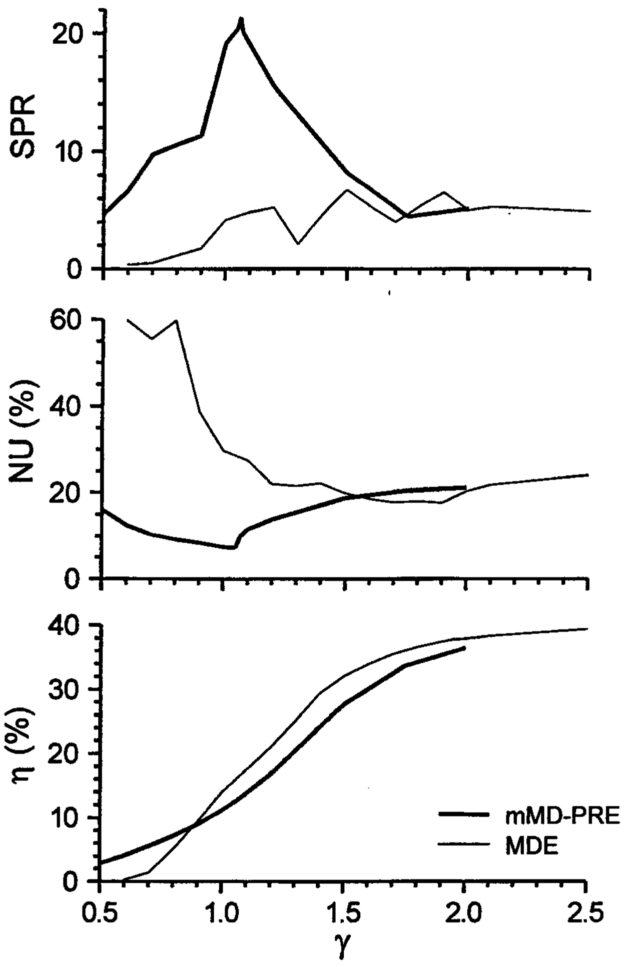


Fig. 4. Performance of MDE and mMD-PRE as a function of the magnitude scaling parameter for encoding the complex-valued function.

NU than does the MDE. These trends and observations are in agreement with those reported in Ref. 12 for various multiphase SLM's. The trends are further brought out in Table 1, which reports the performance for each algorithm when NU is minimum and when SPR is maximum. The table also compares these results with MDE for $\gamma = \infty$ (i.e., the Krackhardt *et al.* biamplitude design). Clearly, the fidelity is much improved by a trade-off of the diffraction efficiency.

Figures 3(a) and 3(b) show a portion of the computer-simulated on-axis diffraction pattern for the MD-PRE and MDE designs reported in Table 1 for minimum NU. Although speckle noise is evident throughout the image of the MD-PRE encoding, there are no significant noise spikes. In Fig. 3(b) most of the background area is pure black; however, there are a number of noise spikes that are quite bright and evident. These noise spikes are due to the inherent nonlinearity of mapping from the continuous real-valued function to the three-valued quantized modulator. The systematic method of mapping in MDE induces strong harmonic terms at sum and difference frequencies of the desired modulation. MD-PRE tends to reduce this effect by distributing noise energy more uniformly over the entire diffraction plane.

Figure 4 presents the performance of encoding the fully complex function by MDE and mMD-PRE as a function of γ. The same sort of trends are seen as a function of γ and in comparing MDE with mMD-PRE as were seen in comparing MDE with MD-PRE. The performance metrics for each algorithm are reported in Table 2. The minimum NU metrics are reported, since they differ only slightly with the maximum SPR design.

The simulated diffraction patterns for the minimum NU design mMD-PRE and MDE designs are given in Figs. 3(c) and 3(d), respectively. Because the SLM can produce only real values and the design produces an off-axis reconstruction, there is a mirror symmetry in both diffraction patterns. Again mMD-PRE has higher SPR than MDE, owing to its distributing the noise over the entire diffraction plane. Comparing Fig. 4 with Fig. 2, we see that the SPR and the diffraction efficiency are substantially smaller in Fig. 4. This is primarily a result of the energy being divided between the desired and the mirror order. The reduction in SPR is also evident when we compare Figs. 3(a) and 3(b) with Figs. 3(c) and 3(d), where the background noise is more evident. However, even with much reduced diffraction efficiency, good performance is possible. As long as the mirror image is acceptable, it appears that a ternary SLM can do a good job of multispot beam steering. Much better performance would be possible with traditional diffractive optical element design approaches. This would involve reoptimizing the design even if the spot array were simply steered to a new location. Such operations would be extremely cumbersome and would limit the adaptivity of many real-time SLM-based systems.

Table 2. Performance for Encoding the Complex-Valued Function

Encoding	γ	η (%)	SPR	NU (%)
mMD-PRE[a]	1.05	12	20.4	7.2
MDE[a]	1.90	37	6.5	17.6
MDE	∞	40	5.3	26.1

[a] Minimum NU.

5. Conclusions

In this paper we extend and reinforce the results originally made in Ref. 12, using quantized phase SLM's. For purposes of producing a Fourier-transform hologram from a desired fully complex-valued function the most faithful encoding method (as measured by SPR and NU) is modified-minimum-distance-pseudorandom encoding (mMD-PRE). Though there is a mirror image for off-axis reconstructions, the real-valued ternary SLM can represent complex-valued functions with good fidelity and moderate diffraction efficiency by use of the mMD-PRE algorithm. The ability to represent complex values on SLM's of such extremely limited modulation is especially useful for reducing the time and cost of prototyping SLM's and SLM-based systems.

This study was supported by Office of Naval Research grant N00014-96-1-1296 and NASA cooperative agreement NCC5-222.

References

1. L. G. Neto, D. Roberge, and Y. Sheng, "Full-range, continuous, complex modulation by the use of two coupled-mode liquid-crystal televisions," Appl. Opt. **35,** 4567–4576 (1996).
2. D. C. Chu, J. R. Fienup, and J. W. Goodman, "Multiemulsion on-axis computer generated hologram," Appl. Opt. **12,** 1386–1388 (1973).
3. B. R. Brown and A. W. Lohmann, "Complex spatial filter," Appl. Opt. **5,** 967 (1966).
4. R. W. Cohn and L. G. Hassebrook, "Representations of fully complex functions on real-time spatial light modulators," in *Optical Information Processing*, F. T. S. Yu and S. Jutamulia, eds. (Cambridge University, Cambridge, UK, 1998), Chap. 15, pp. 396–432.
5. J. N. Mait, "Diffractive beauty," Opt. Photon. News **52,** 21–25 (1998).
6. W.-H. Lee, "Computer-generated holograms: techniques and applications," in *Progress in Optics*, E. Wolf, ed. (Elsevier, Amsterdam, 1978), Vol. 16, pp. 119–231.
7. W. J. Dallas, "Computer-generated holograms," in *The Computer in Optical Research*, B. R. Frieden, ed. (Springer-Verlag, Berlin, 1980), Chap. 6, pp. 291–366.
8. R. W. Cohn, "Pseudorandom encoding of fully complex functions onto amplitude coupled phase modulators," J. Opt. Soc. Am. A **15,** 868–883 (1998).
9. R. D. Juday, "Optimal realizable filters and the minimum Euclidean distance principle," Appl. Opt. **32,** 5100–5111 (1993).
10. R. W. Cohn and Markus Duelli, "Ternary pseudorandom encoding of Fourier transform holograms," J. Opt. Soc. Am. A **16,** 71–84 (1999).
11. L. G. Hassebrook, M. E. Lhamon, R. C. Daley, R. W. Cohn, and M. Liang, "Random phase encoding of composite fully complex filters," Opt. Lett. **21,** 272–274 (1996).
12. M. Duelli, M. Reece, and R. W. Cohn are preparing a manuscript to be called "Modified minimum distance criterion for blended random and nonrandom encoding."
13. B. A. Kast, M. K. Giles, S. D. Lindell, and D. L. Flannery, "Implementation of ternary phase amplitude filters using a magnetooptic spatial light modulator," Appl. Opt. **28,** 1044–1046 (1989).
14. *VLSI Spatial Light Modulators*, Operations Manual for 128 × 128 analog SLM, revision 1.6 (Boulder Nonlinear Systems Inc., Lafayette, Colo., 1998).
15. R. W. Cohn and M. Liang, "Approximating fully complex spatial modulation with pseudorandom phase-only modulation," Appl. Opt. **33,** 4406–4415 (1994).
16. R. W. Cohn and M. Liang, "Pseudorandom phase-only encoding of real-time spatial light modulators," Appl. Opt. **35,** 2488–2498 (1996).
17. R. W. Cohn, "Analyzing the encoding range of amplitude-phase coupled spatial light modulators," Opt. Eng. **38,** 361–367 (1999).
18. U. Krackhardt, J. N. Mait, and N. Streibl, "Upper bound on the diffraction efficiency of phase-only fanout elements," Appl. Opt. **31,** 27–37 (1992).

Analyzing the encoding range of amplitude-phase coupled spatial light modulators

Robert W. Cohn, MEMBER SPIE
University of Louisville
The ElectroOptics Research Institute
Louisville, Kentucky 40292
E-mail: rwcohn01@ulkyvm.louisville.edu

Abstract. Most spatial light modulators (SLMs) are limited in that they cannot produce arbitrary complex modulations. Because phase and amplitude are usually coupled, it is difficult to computer design appropriate modulation patterns fast enough for the real-time applications for which SLMs are suited. Dramatic computational speedups can be achieved by using encoding algorithms that directly translate desired complex values into values that the modulator can produce. For coherently illuminated SLMs in a Fourier transform arrangement, pseudorandom encoding can be used. Each SLM pixel is programmed in sequence by selecting a single value of pixel modulation from a random distribution having an average that is identical to the desired fully complex modulation. While the method approximates fairly arbitrary complex modulations, there are always some complex values that are outside the encoding range for each SLM coupling characteristic and for each specific pseudorandom algorithm. Using the binary random distribution leads to methods of evaluating and geometrically interpreting the encoding range. Evaluations are presented of achieving fully complex encoding with SLMs that produce less than 2π of phase shift, identifying an infinite set of encoding algorithms that encode the same value, identification of the maximum encoding range, and geometric interpretation of encoding errors. [S0091-3286(99)01702-X]

Subject terms: spatial light modulators; computer generated holography; complex-valued encoding; statistical optics.

Paper 980184 received May 5, 1998; revised manuscript received Sep. 14, 1998; accepted for publication Sep. 24, 1998. This paper is a revision of a paper presented at the SPIE Conference on Spatial Light Modulators, January 29–30, 1998, San Jose, CA. The paper presented there appears (unrefereed) in Proc. SPIE Vol. 3297 (1998).

1 Introduction

The encoding of fully complex functions onto computer generated Fourier transform holograms was first introduced by Brown and Lohmann.[1,2] The method and other subsequent fully complex encoding methods, reviewed in Refs. 3 to 5, provide the ability to transparently specify desired far-field diffraction patterns in terms of Fourier transform identities, tables and other well known relationships, many that are known in closed form. An example of a system application is the potential of using phase-only spatial light modulators (SLMs) to produce and steer multiple spots in arbitrary directions independent of each other in real time.[6] The ability to directly encode and represent desired complex valued modulation provides the numerical efficiency required to design a continuous stream of modulations in real time. Other real-time applications that can benefit from complex-valued representations are considered in Ref. 5. Additional advantages of complex valued representations in terms of the fidelity of computer generated holograms (CGHs) and diffractive optical elements were also considered by Kettunen et al.[7] Certainly the recognition of these various advantages has also spurred the development and demonstration of fully complex modulators.[8–11] No fully complex SLM is commercially available, however, and the most recent demonstrations of fully complex SLMs require the use of multiple SLMs. For these reasons, complex-valued representations continue to be of interest.

Most of the early CGH design methods treat multiple pixels as a single group to realize a single complex value. By grouping pixels, the space-bandwidth product of the signal is necessarily less than that of the CGH. Therefore the useable bandwidth of the reconstruction is limited to only a fraction of the entire bandwidth set by the pixel sample spacing. This is especially important today when electrically addressed SLMs are relatively expensive and consist of a small number of pixels compared to fixed pattern CGHs and diffractive optics.

One early technique that does use a single pixel to represent a complex value is the original kinoform, in which the magnitudes of each complex value are set to unity.[12] Due to noise and inaccuracies in the reconstruction,[7] however, most phase-only CGHs now are designed using various numerically intensive global search algorithms.[13–18] In some real-time systems, the filter design may need to be done on-line, which may not enable global searches to be performed. Other single pixel methods are the minimum

Euclidean distance (MED) methods for matched filter[19] and CGH (Ref. 20) design. MED optimizes a performance metric as a function of a complex-valued factor that scales all the desired complex values. The performance metric is calculated for each value of the scale factor. Once the optimum scale factor is found, each desired complex value is mapped to the closest realizable value on the modulation characteristic. This two parameter search requires considerably less computation to perform than the global searches.

MED and related studies[21,22] are important in that they describe filter design methods in such ways that the methods apply to a wide variety of modulation characteristics. This recognizes the fact that current electrically and optically addressed SLMs have widely varied modulation characteristics that are usually not accurately described as being pure phase or amplitude modulators. Instead, these SLMs exhibit various degrees of coupling between amplitude and phase, as reported in a number of papers on SLM modulation characteristics.[23–26]

Pseudorandom encoding,[27] the subject of this paper, is a single pixel method that has been primarily applied to CGH design. The procedure for mapping a desired complex value to a realizable pixel modulation is a noniterative and numerically efficient operation. When the CGH is illuminated by a uniform plane wave, the far-field diffraction pattern approximates the desired reconstruction in an average sense. Superimposed on the desired reconstruction is a white noise pattern that covers the entire reconstruction plane. The energy in the noise is equivalent to the errors between the desired complex values and the realized values. By spreading the noise over the full extent of the reconstruction plane, the noise level is, on average, the lowest possible for a fixed level of error energy. This can be compared[28] with error diffusion methods for CGHs. The reconstruction from the error diffused hologram also produces a noise cloud, but the noise cloud and desired reconstruction appear in different regions of the reconstruction plane. Thus, unlike error diffusion, pseudorandom encoding enables the desired reconstruction to be formed anywhere in the full extent of the reconstruction plane. The noise level from pseudorandom encoding a particular signal still might be unacceptably high, however, there are simple calculations that can be performed prior to encoding that measure the noise,[27,29] and many pseudorandom encoded designs with negligible noise backgrounds have been reported to date.[27,29–31]

Originally pseudorandom encoding was introduced[27] for phase-only SLMs, and then ways to generalize this method to amplitude-phase coupled modulators were considered.[31] A useful result from this study was the development of a closed form encoding algorithm in which the values of any given coupled modulation characteristic could be explicitly placed in the formula.[31] However, numerical evaluations show that some complex values could not be encoded by this method. These observations led to more fundamental analyses of the properties of pseudorandom encoding, including the encoding range and the amount of error signal produced by encoding. The greatest progress was made by using binary statistics that, in addition to numerical ease, provide useful geometrical interpretations of various properties of the pseudorandom encoding methods. This paper specifically describes this analysis technique and presents examples that, in addition to illustrating the analysis technique, are used to identify properties of pseudorandom encoding that were not previously known. Sec. 2 reviews the mathematics of pseudorandom encoding and specializes the problem for the use of binary statistics. Then in Sec. 3, these tools are applied to evaluating the encoding range and encoding errors for SLMs having a variety of modulation characteristics.

2 Pseudorandom Encoding of Fully Complex Functions

2.1 *General Description of Pseudorandom Encoding*

All pseudorandom encoding algorithms specify the modulation of any given pixel in terms of a user specified random variable. The statistical properties of the random variable are selected in such a way that the expected value, or average, of the random modulation is identical to the desired, but unobtainable, fully complex value. The desired complex-valued modulation is written $\mathbf{a}_c=(a_c,\psi_c)$ and the resulting modulation by the SLM is $\mathbf{a}=(a,\psi)$, where the ordered pairs are the polar representations of the complex quantities. Complex quantities are indicated by bold type. The pseudorandom encoding design statement is, in general, to find a value of the ensemble average

$$\langle\mathbf{a}\rangle=\int \mathbf{a}p(\mathbf{a})d\mathbf{a}, \tag{1}$$

of the random variable $\mathbf{a}$ such that $\langle\mathbf{a}\rangle=\mathbf{a}_c$. The statistical properties of $\mathbf{a}$ are determined by its probability density function $p(\mathbf{a})$. The probability density function (pdf) is *selected* to ensure that the expected value of $\mathbf{a}$ and the desired complex value are identical. This *selection* of a pdf corresponds to solving the integral equation, Eq. (1) for $p(\mathbf{a})$. The solution is not unique since the integral in Eq. (1) is a projection from the multidimensional space of $\mathbf{a}$ into a single value $\langle\mathbf{a}\rangle$. After an appropriate density function is determined, the desired complex value $\mathbf{a}_c$ is encoded by drawing a single value of $\mathbf{a}$ from a random distribution having the density function $p(\mathbf{a})$. Since the value of $\mathbf{a}$ is found deterministically by computer, rather than from a random process occurring in nature, the procedure has been termed pseudorandom encoding.

This general pseudorandom encoding prescription is applied to each pixel in sequence to encode the desired spatially varying complex modulations $\mathbf{a}_c$. Using i as the spatial coordinate, the spatial samples of the desired complex modulation, the density function and the random modulation can be written as $\mathbf{a}_{ci}$, $p_i(\mathbf{a}_i)$ and $\mathbf{a}_i$. (This indexing scheme can be conveniently applied to 1-D or 2-D arrays and it is not restricted to equally spaced samples.)

The far-field diffraction pattern of the encoded modulation $\mathbf{a}_i$ approximates the desired diffraction pattern. This can be seen by comparing the intensity of the desired far-field diffraction pattern with the ensemble average diffraction pattern that would result from the encoded modulation. The intensity pattern of the desired diffraction pattern from an N sample fully complex SLM is

$$I_c(f_x)=\left|\sum_{i=1}^{N}\mathbf{A}_{ci}\right|^2=\left|\mathcal{F}\left(\sum_{i=1}^{N}\mathbf{a}_{ci}\right)\right|^2, \tag{2}$$

where $\mathcal{F}\{\cdot\}$ is the Fourier transform operator, $\mathbf{A}_{ci}(f_x)$ is the Fourier transform of the transmittance of the i'th pixel of the SLM, and f_x is the spatial coordinate across the Fourier plane. The expected intensity of the diffraction pattern from the encoded modulation was derived for the condition that the random variable $\mathbf{a}_i$ for the i'th pixel is statistically independent of $\mathbf{a}_j$ for all j not equal to i. Under the pseudorandom design condition $\langle\mathbf{a}_i\rangle=\mathbf{a}_{ci}$ the ensemble average pattern is expressed[27,29]

$$\langle I(f_x)\rangle=I_c(f_x)+\sum_{i=1}^{N}(\langle|\mathbf{A}_i|^2\rangle-|\mathbf{A}_{ci}|^2), \tag{3}$$

where $\mathbf{A}_i(f_x)$ is the Fourier transform of $\mathbf{a}_i$. The expected intensity consists of two terms. The first term is the desired diffraction pattern from Eq. (2). The second term, the N term summation, represents the average level of background (i.e., speckle) noise that is produced as a result of the randomness of the modulation. It is the error signal referred to in the introduction. For the case of pixels that are modeled as pointlike apertures, the average background noise is of constant intensity for all frequencies f_x (i.e., it is white.)

2.2 *Encoding Error Defined*

Eq. (3) identifies individually the noise contribution of each pixel. Therefore insight can be gained by evaluating the noise contribution in the modulation plane. Under the assumption that the pixels are infinitesimally wide apertures, the inverse Fourier transform of the noise from a single SLM pixel [i.e., a single term from the summation in Eq. (3)] gives the encoding error

$$\varepsilon=\langle|\mathbf{a}|^2\rangle-|\mathbf{a}_c|^2, \tag{4}$$

where the subscript has been dropped to simplify presentation. (If the pixels are finite width, then an autocorrelation of the pixel aperture function would also be included in the formula.[27] This term is dropped because it adds no essential insight to the current discussion.) In the next subsection Eqs. (1) and (4) are specialized for the case where $\mathbf{a}$ is a binary random variable.

2.3 *Pseudorandom Encoding with the Binary Distribution and Geometric Interpretation*

The probability density function for the binary distribution is

$$p(\mathbf{a})=d\delta(\mathbf{a}-\mathbf{a}_1)+(1-d)\delta(\mathbf{a}-\mathbf{a}_2), \tag{5}$$

where $\delta(\cdot)$ is the Dirac delta function, $\mathbf{a}_1$ and $\mathbf{a}_2$ are a pair of complex values from the modulation characteristic and d and $1-d$ are the probabilities of selecting $\mathbf{a}_1$ and $\mathbf{a}_2$. Since d is a probability, its value is between 1 and 0. Using the binary density function in Eq. (1) gives an expression for the effective complex amplitude of

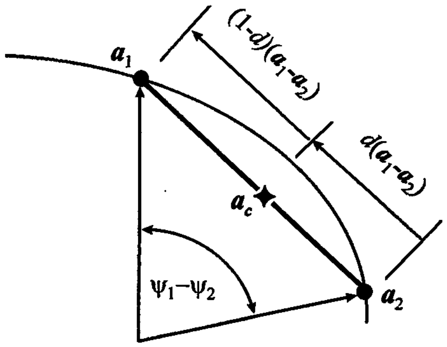


Fig. 1 Geometric relationships for pseudorandom encoding the desired value $\mathbf{a}_c$ using random binary selection of modulator values $\mathbf{a}_1$ and $\mathbf{a}_2$.

$$\langle\mathbf{a}\rangle=d\mathbf{a}_1+(1-d)\mathbf{a}_2. \tag{6}$$

Eq. (6) is recognized as the expression for a line as a function of the variable d. For $d=1$, $\mathbf{a}_1$ is encoded, for $d=0$, $\mathbf{a}_2$ is encoded and for values of d between 1 and 0, any value lying on the line segment between $\mathbf{a}_1$ and $\mathbf{a}_2$ can be encoded. Therefore, the *encoding range* of pseudorandom binary encoding is the line segment that connects $\mathbf{a}_1$ to $\mathbf{a}_2$ as illustrated in Fig. 1.

2.3.1 *The binary encoding formula*

For a given value of d the desired complex value $\mathbf{a}_c(d)=\langle\mathbf{a}\rangle$ is represented (i.e., encoded) by a single randomly selected value

$$\begin{aligned}\mathbf{a}&=\mathbf{a}_1 \quad \text{if} \quad 0\leqslant s\leqslant d,\\ \mathbf{a}&=\mathbf{a}_2 \quad \text{if} \quad d<s\leqslant 1,\end{aligned} \tag{7}$$

where s is a uniformly distributed random number between 0 and 1.

2.3.2 *Binary encoding error*

Evaluating Eq. (4) using Eqs. (5) and (6) and the expectation

$$\langle|\mathbf{a}|^2\rangle=d|\mathbf{a}_1|^2+(1-d)|\mathbf{a}_2|^2, \tag{8}$$

the encoding error can be written

$$\varepsilon=d(1-d)[a_1^2+a_2^2-2a_1a_2\cos(\psi_1-\psi_2)]. \tag{9}$$

The error is written in terms of the magnitudes and phases of $\mathbf{a}_i=(a_i,\psi_i)$ to show that the term in the brackets is the familiar formula for the "law of cosines," which gives the squared magnitude of the line segment $\mathbf{a}_1-\mathbf{a}_2$. Thus the encoding error can be written

$$\varepsilon=d(1-d)|\mathbf{a}_1-\mathbf{a}_2|^2. \tag{10}$$

Using the main premise of pseudorandom encoding that $\mathbf{a}_c\equiv\langle\mathbf{a}\rangle$, Eq. (6) can be rearranged to make the two relationships evident

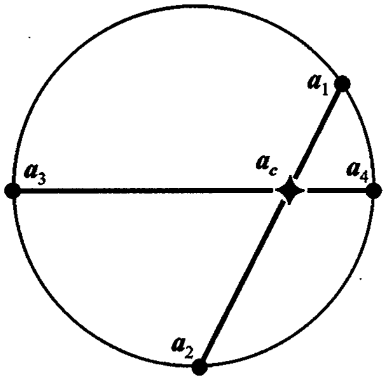


Fig. 2 Multiple possible pairs of modulation values (joined by chords) that can pseudorandom encode the desired value $\mathbf{a}_c$ for a circular modulation characteristic. This construction shows that there are an infinite number of possible binary pairs that encode $\mathbf{a}_c$.

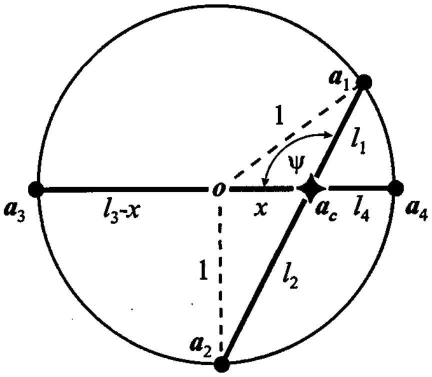


Fig. 3 Geometric relationships used to prove that for a circular modulation characteristic, the encoding error is identical for encoding with any pair of points from the modulation characteristic that are collinear with the desired value $\mathbf{a}_c$.

$$\mathbf{a}_c - \mathbf{a}_2 = d(\mathbf{a}_1 - \mathbf{a}_2),$$
$$\mathbf{a}_1 - \mathbf{a}_c = (1-d)(\mathbf{a}_1 - \mathbf{a}_2). \tag{11}$$

These lengths are indicated on Fig. 1. Using Eq. (11) in Eq. (10) shows that binary pseudorandom encoding error can be expressed as

$$\varepsilon = |\mathbf{a}_1 - \mathbf{a}_c||\mathbf{a}_c - \mathbf{a}_2|. \tag{12}$$

Written in this form, the encoding error can be directly interpreted as the product of the lengths of the line segments $\mathbf{a}_1$ to $\mathbf{a}_c$ and $\mathbf{a}_c$ to $\mathbf{a}_2$.

This section has (1) identified [following Eq. (6)] that the pseudorandom encoding range for any pair of complex valued points is the line segment connecting those two points and (2) derived [Eq. (12)] that the encoding error is equal to the product of the lengths of the two line segments that connect $\mathbf{a}_1$ to $\mathbf{a}_c$ and $\mathbf{a}_2$ to $\mathbf{a}_c$. These two basic results provide a useful tool for evaluating and understanding the encoding properties of various coupled SLM characteristics and, ultimately, developing new pseudorandom encoding algorithms for specific modulation characteristics. Their application to the analysis of a variety of coupled modulation characteristics is illustrated by the examples that follow in Sec. 3.

3 Illustrative Analyses of Pseudorandom Encoding on Coupled SLMs

3.1 *Phase-Only and Circular Modulation Characteristics*

Figure 2 shows a circular modulation characteristic on the complex plane. This is more general than phase-only because the center of the modulation characteristic (or curve) is not necessarily located at zero. Off centered curves are typical of birefringent liquid crystal SLMs used with a polarizer. As the polarizer is rotated away from the extraordinary axis, the center of the modulation characteristic moves away from the origin in the complex plane. Thus this characteristic can be viewed as coupled in amplitude and phase. However, it is not necessary to write the relationship between amplitude and phase for the evaluations to be considered here.

Figure 2 illustrates two interesting properties of pseudorandom encoding for circular modulation characteristics. First, it can be seen by repeated plottings of Eq. (6) for different values of $\mathbf{a}_i$ and $\mathbf{a}_2$ that any complex value inside the circular characteristic can be encoded by choosing a pair of points from the characteristic that are colinear with the desired complex value $\mathbf{a}_c$. Second, it can be seen that since probability d of selecting one endpoint can never exceed unity, no complex values outside the characteristic can be pseudorandom encoded. Third, it can be seen that there are multiple pairs of points that can encode the same value $\mathbf{a}_c$. Obviously there are an infinite number of solutions. This nonuniqueness of pseudorandom encoding was described in general in the text following Eq. (1).

If there are multiple solutions possible, is there one particular one that produces the least encoding error? It turns out that for circular characteristics, the encoding error is identical for all possible solutions. This can be proved using the geometric constructions in Fig. 3. For this statement to be true then according to Eq. (12)

$$\varepsilon_c = l_1 l_2 = l_3 l_4, \tag{13}$$

where l_i are the distances from the points $\mathbf{a}_i$ to $\mathbf{a}_c$. The line segment $\mathbf{a}_1$ to $\mathbf{a}_2$ in Fig. 3 passes through the center of the circle. The radius of the circle can be assumed to be unity with no loss in generality. Also defining $x = 1 - l_3$ as the distance from the center to $\mathbf{a}_c$ leads to a straightforward derivation. The law of cosines gives the expressions

$$l_1^2 - 2xl_1 \cos\psi - (1-x^2) = 0,$$
$$l_2^2 + 2xl_2 \cos\psi - (1-x^2) = 0, \tag{14}$$

where $-\cos\psi = \cos(\pi - \psi)$ has been used to eliminate the complementary angle $\pi - \psi$ from the second expression in Eq. (14). The positive roots to the quadratic equations are

$$l_1 = x\cos\psi + (x^2\cos^2\psi + 1 - x^2)^{1/2},$$

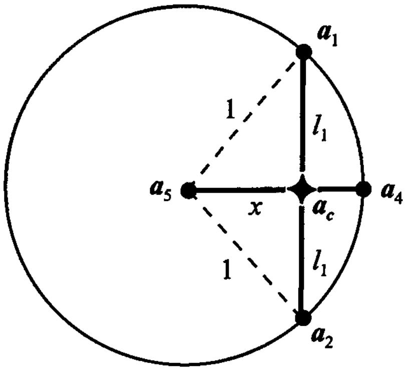


Fig. 4 Constructions used in the comparison of the encoding error for biradial and circular modulation characteristics. We show in the text that when encoding the desired value $\mathbf{a}_c$, the error is always lower if the biradial modulation values $\mathbf{a}_4$ and $\mathbf{a}_5$ are used rather than the values $\mathbf{a}_1$ and $\mathbf{a}_2$ from the single radius portion of the modulation characteristic.

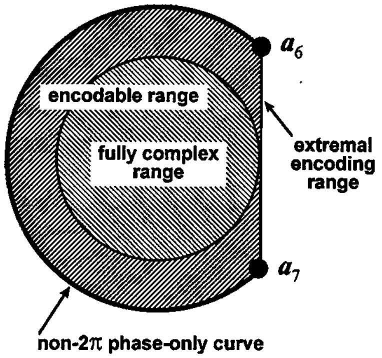


Fig. 5 Encoding range (shaded) for a non-2π phase-only SLM (thick curve). The thin line represents the values that can be pseudorandom encoded using the endpoints $\mathbf{a}_6$ and $\mathbf{a}_7$ of the modulation characteristic. The fully complex range is the largest circular region surrounding the origin. Therefore the non-2π modulator can represent a fully complex modulator if the desired complex-valued modulation are scaled so that their values fit within the fully complex region.

$$l_2 = -x \cos \psi + (x^2 \cos^2 \psi + 1 - x^2)^{1/2}. \tag{15}$$

The product of the two lengths then simplifies to

$$\varepsilon_c = l_1 l_2 = (1+x)(1-x), \tag{16}$$

which is independent of angle ψ and which is seen by inspection of Fig. 3 to be identical to $l_3 l_4$.

3.2 *Biradial Circular Modulation Characteristics*

Figure 4 shows a circular SLM characteristic that, in addition to the Fig. 2 characteristic, also can produce a modulation state at the center of the circle. For characteristics centered on the origin, this would be a phase-only SLM that also contains a 0 state. Using the geometric construction in Fig. 4, we can see that the encoding error produced by encoding with the points $\mathbf{a}_5$ and $\mathbf{a}_4$ is

$$\varepsilon_b = l_5 l_4 = x(1-x). \tag{17}$$

From the result in Eq. (16) or from the geometry in Fig. 4 we can see that encoding with the central point always produces less error than does biamplitude encoding with a phase-only SLM. For a unit radius circular characteristic $x=d$ and the ratio of the two types of error reduces to

$$\frac{\varepsilon_c}{\varepsilon_b} = 1 + \frac{1}{d}. \tag{18}$$

Thus the encoding errors are always lower for the biradial SLM and substantially better when the desired complex values are close to zero.

3.3 *Non-2π SLMs, Discrete SLMs and Convoluted Modulation Characteristics*

Figure 5 shows a phase-only SLM that does not produce a full 2π of phase modulation. The curve starts with point $\mathbf{a}_6$ and ends with point $\mathbf{a}_7$. Obviously, any value on a line between these two points can be pseudorandom encoded. Also line segments can be drawn that fill in the interior of the region bounded by the modulation curve and the extremal line segment $\mathbf{a}_6$ to $\mathbf{a}_7$. The interior region (shown by both shading patterns) forms a convex set. The key result of this analysis is that by appropriately scaling the desired complex values (to fit within the circular, fully complex region in Fig. 5) it would be possible to pseudorandom encode any fully complex function with the non-2π phase-only characteristic.

Figure 6 shows a convoluted characteristic. Some complex values that in a sense are *outside* the modulation characteristic can also be encoded by binary pseudorandom encoding. There are three extremal line segments in Fig. 6 that, together with the convex portions of the modulation curve, define the boundary on the convex set of encodable values. Using binary encoding analysis to evaluate the encoding range shows that the encoding range can be significantly larger than one might at first assume.

The encoding range of a discrete modulation characteristic is evaluated in Fig. 7. Here only values on the six line

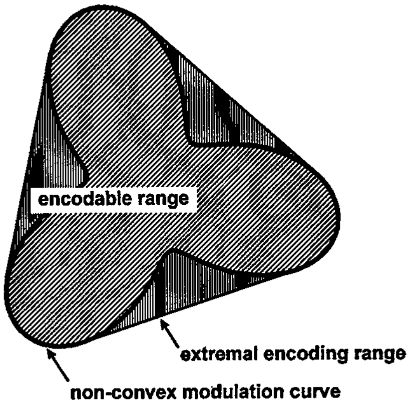


Fig. 6 Encoding range for a nonconvex modulation characteristic (thick curve). The three thin lines together with the convex portions of the SLM bound the convex region that can be pseudorandom encoded. The two shading patterns are used to distinguish the nonconvex region inside the modulation characteristic from the additional range that the analysis using the properties of binary statistics has identified as being encodable.

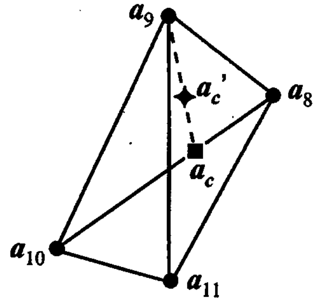


Fig. 7 Encoding range for a four-value discrete modulation characteristic. Binary pseudorandom encoding only encodes values on the six line segments. However, any other value (e.g., $\mathbf{a}_c'$) in the convex region bounded by the lines connecting $\mathbf{a}_8$, $\mathbf{a}_9$, $\mathbf{a}_{10}$, $\mathbf{a}_{11}$, and $\mathbf{a}_8$ could be encoded by a combination of binary encoding algorithms.

segments connecting the four realizable points can be pseudorandom encoded *using binary distributions.* However, other random distributions can be used to encode the entire region surrounded by curve $\mathbf{a}_8,\mathbf{a}_9,\mathbf{a}_{10},\mathbf{a}_{11},\mathbf{a}_8$. One approach for doing this is to build up more involved distributions out of combinations of binary distributions. The dashed line in Fig. 7 represents the set of values that could be encoded between $\mathbf{a}_9$ and $\mathbf{a}_c$ if the value $\mathbf{a}_c$ were part of the modulation characteristic. However the value $\mathbf{a}_c$ is the result of binary pseudorandom encoding using the values $\mathbf{a}_8$ and $\mathbf{a}_9$, and it is not part of the characteristic. Nonetheless it is possible to randomly select between $\mathbf{a}_9$ and $\mathbf{a}_c$ so that the desired value $\mathbf{a}_c'$ can be realized on average. This two step encoding formula was evaluated as being equivalent to a single encoding formula using a ternary distribution that selects between the points $\mathbf{a}_8$, $\mathbf{a}_9$, and $\mathbf{a}_{10}$. A derivation of such a formula (though not applied to discrete SLMs) is given in Ref. 31. From this analysis, it becomes apparent that by varying the value of $\mathbf{a}_c$ it is possible to encode the entire region interior to $\mathbf{a}_8,\mathbf{a}_9,\mathbf{a}_{10},\mathbf{a}_8$ using a ternary encoding formula. The region interior to $\mathbf{a}_8,\mathbf{a}_{10},\mathbf{a}_{11},\mathbf{a}_8$ can be pseudorandom encoded in a similar manner.

4 Summary and Conclusions

The use of binary statistics leads to extremely simple pseudorandom encoding formulas. Perhaps the greatest value of using the binary distribution is that it provides significant insight and even a graphical interpretation of the operation and performance of pseudorandom encoding. Evaluations have been presented for a variety of amplitude-phase coupled SLM characteristics. In each case a convex region was identified by simple graphical constructions. An extremely simple formula for determining the encoding error was presented. Applying it to circular SLM characteristics showed there to be no advantage as to which pair of modulation values are used in the encoding formula. For noncircular SLM characteristics, however, significant reductions in encoding error are possible. We also found that phase-only SLMs that produce phase modulations greater than π but less than 2π can also be pseudorandomly encoded with fully complex functions. Also discussed were ways to build up more complicated functions out of combinations of binary distributions. This was used to demonstrate that discrete modulation characteristics can represent fully complex functions over a continuous region in the complex plane. This result is especially significant in light of the digital addressing of SLMs and the small number of phase steps used in most multilevel binary diffractive optics. Therefore, this analysis method proves to be a useful tool that can accelerate the development and broaden the applicability of pseudorandom encoding algorithms to a wide variety of amplitude-phase coupled modulator characteristics.

Acknowledgments

This study was supported by the National Aeronautics and Space Administration (NASA) cooperative agreement No. NCC5-222 and Office of Naval Research (ONR) grant No. N00014-96-1-1296.

References

1. B. R. Brown and A. W. Lohmann, "Complex spatial filter," *Appl. Opt.* **5**, 967–969 (1966).
2. B. R. Brown and A. W. Lohmann, "Computer-generated binary holograms," *IBM J. Res. Dev.* **13**, 160–168 (1969).
3. W. J. Dallas, "Computer-generated holograms," Chap. 6 in *The Computer in Optical Research*, B. R. Frieden, Ed., pp. 291–366, Springer, Berlin (1980).
4. O. Bryngdahl and F. Wyrowski, "Digital holography—computer-generated holograms," in *Progress in Optics XXVIII*, E. Wolf, Ed., pp. 1–86, Elsevier, Amsterdam (1990).
5. R. W. Cohn and L. G. Hassebrook, "Representations of fully complex functions on real-time spatial light modulators," Chap. 15, in *Optical Information Processing*, F. T. S. Yu and S. Jutamulia, Eds., pp. 396–432, Cambridge University Press, Cambridge (1998).
6. R. W. Cohn, "Real-time multispot beam steering with electrically controlled spatial light modulators," in *Optical Scanning Systems: Design and Applications*, L. Beiser and S. F. Sagan, Eds., *Proc. SPIE* **3131**, 145–155 (July 1997).
7. V. Kettunen, P. Vahimaa, and J. Turunen, "Zeroth-order coding of complex amplitude in two dimensions," *J. Opt. Soc. Am. A* **14**(4), 808–815 (1997).
8. J. M. Florence and R. D. Juday, "Full complex spatial filtering with a phase mostly DMD," *Proc. SPIE* **1558**, 487–498 (1991).
9. R. D. Juday and J. M. Florence, "Full complex modulation with two one-parameter SLMs," *Proc. SPIE* **1558**, 499–504 (1991).
10. D. A. Gregory, J. C. Kirsch and E. C. Tam, "Full complex modulation using liquid-crystal televisions," *Appl. Opt.* **31**(2), 163–165 (1992).
11. L. G. Neto, D. Roberge and Y. Sheng, "Full-range, continuous, complex modulation by the use of two coupled-mode liquid-crystal televisions," *Appl. Opt.* **35**(23), 4567–4576 (1996).
12. L. B. Lesem, P. M. Hirsch and J. A. Jordon, Jr., "The kinoform: a new wavefront reconstruction device," *IBM J. Res. Dev.* **13**, 150–155 (1969).
13. R. W. Gerchberg and W. O. Saxton, "Practical algorithm for the determination of phase from image and diffraction plane pictures," *Optik (Stuttgart)* **35**(2), 237–250 (1972).
14. N. C. Gallagher and B. Liu, "Method for computing kinoforms that reduces image reconstruction error," *Appl. Opt.* **12**(10), 2328–2335 (1973).
15. H. Stark, W. C. Catino, and J. L. LoCicero, "Optical phase-mask design using generalized projections," *J. Opt. Soc. Am. A* **8**(3), 566–571 (Mar. 1991).
16. M. P. Dames, R. J. Dowling, P. McKee, and D. Wood, "Efficient optical elements to generate intensity weighted spot arrays: design and fabrication," *Appl. Opt.* **30**(19), 2685–2691 (1991).
17. E. G. Johnson, M. A. Abushagur, "Microgenetic-algorithm optimization methods applied to dielectric gratings," *J. Opt. Soc. Am. A* **12**(5), 1152–1160 (1995).
18. J. N. Mait, "Understanding diffractive optic design in the scalar domain," *J. Opt. Soc. Am. A* **12**(10), 2145–2158 (1995).
19. R. D. Juday, "Optimal realizable filters and the minimum Euclidean distance principle," *Appl. Opt.* **32**(26), 5100–5111 (1993).
20. R. D. Juday and J. Knopp, "HOLOMED—an algorithm for computer generated holograms," *Proc. SPIE* **2752**, 162–172 (1996).
21. M. W. Farn and J. W. Goodman, "Optimal maximum correlation filter for arbitrarily constrained devices," *Appl. Opt.* **28**(15), 3362–3366 (1989).
22. R. D. Juday, "Correlation with a spatial light modulator having phase

and amplitude cross coupling," *Appl. Opt.* **28**(22), 4865–4869 (1989).
23. K. Lu and B. E. A. Saleh, "Theory and design of the liquid crystal TV as an optical spatial phase modulator," *Opt. Eng.* **29**(3), 241–246 (1990).
24. C. Zeile and E. Luder, "Complex transmission of liquid crystal spatial light modulators in optical signal processing applications," *Proc. SPIE* **1911**, 195–206 (1993).
25. J. L. Pezzaniti and R. A. Chipman, "Phase-only modulation of twisted nematic liquid-crystal TV by use of the eigenpolarization states," *Opt. Lett.* **18**(18), 1567–1569 (1993).
26. C. Soutar, S. E. Monroe, Jr., and J. Knopp, "Measurement of the complex transmittance of the Epson liquid crystal television," *Opt. Eng.* **33**(4), 1061–1068 (1994).
27. R. W. Cohn and M. Liang, "Pseudorandom phase-only encoding of real-time spatial light modulators," *Appl. Opt.* **35**(14), 2488–2498 (1996).
28. R. Hauck and O. Bryngdahl, "Computer-generated holograms with pulse-density modulation," *J. Opt. Soc. Am. A* **1**(1), 5–10 (1984).
29. R. W. Cohn and W. Liu, "Pseudorandom encoding of fully complex modulation to bi-amplitude phase modulators," *Optical Society of America Topical Meeting on Diffractive Optics and Micro Optics*, 1996 OSA Technical Digest Series, Vol. 5, pp. 237–240, Boston, MA (1996).
30. R. W. Cohn, A. A. Vasiliev, W. Liu and D. L. Hill, "Fully complex diffractive optics via patterned diffuser arrays," *J. Opt. Soc. Am. A* **14**(5), 1110–1123 (1997).
31. R. W. Cohn, "Pseudorandom encoding of fully complex functions onto amplitude coupled phase modulators," *J. Opt. Soc. Am. A* **15**(4), 868–883 (1998).

Robert W. Cohn is a professor of electrical engineering and director of the ElectroOptics Research Institute at the University of Louisville. He holds his PhD degree from Southern Methodist University and his MS and BS degrees from the University of Kansas, Lawrence, all in electrical engineering. Prior to joining the university, from 1978 to 1989 Cohn was a member of the technical staff of Texas Instruments in Dallas, performing research on spatial light modulators for optical information processing, surface acoustic wave devices and microwave hybrid circuits for electronic signal processing, and tracking algorithms for imaging IR missile seekers. At the University of Louisville he continues research on the application and characterization of spatial light modulators as well as related work on the design, fabrication and measurement of diffractive optics. He is a member of SPIE and OSA and a senior member of IEEE.

Ternary pseudorandom encoding of Fourier transform holograms

Robert W. Cohn and Markus Duelli

The ElectroOptics Research Institute, University of Louisville, Louisville, Kentucky 40292

Received June 1, 1998; revised manuscript received September 10, 1998; accepted September 14, 1998

Pseudorandom encoding is a statistical method for designing Fourier transform holograms by mapping ideal complex-valued modulations onto spatial light modulators that are not fully complex. These algorithms are notable because their computational overhead is low and because the space–bandwidth product of the encoded signal is identical to the number of modulator pixels. All previous pseudorandom-encoding algorithms were developed for analog modulators. A less restrictive algorithm for quantized modulators is derived that permits fully complex ranges to be encoded with as few as three noncollinear modulation values that are separated by more than 180° on the complex plane. © 1999 Optical Society of America [S0740-3232(99)02001-3]

OCIS codes: 230.6120, 090.1760, 030.6600, 070.0070.

1. INTRODUCTION

A. Rationale for Using Pseudorandom Encoding for the Design of Fourier Transform Holograms

The first computer-generated hologram (CGH) solved the problem of representing complex-valued modulations with a binary amplitude-only transmittance.[1] High-quality reconstructions were possible because of the high spatial resolution of large-area plotters (followed by successive photographic reductions). Since then various CGH algorithms have been developed in response to the particular physical properties of the modulating medium—physical properties such as modulating type: amplitude-only, phase-only, or coupled amplitude–phase modulation; modulation levels: continuous or quantized; spatial structure: continuous or discretely sampled; spatial resolution/space–bandwidth product: low to high; and update rate: fixed-pattern to programmable in real time.[2]

CGH algorithms are also shaped by the intended application. For example, for today's fixed-pattern diffractive optic Fourier transform holograms that are replicated *en masse*, there is no major time constraint in employing design algorithms that use numerically intensive optimizations and search strategies. However, if individual custom-designed CGH's are to be used by a large customer base (e.g., a unique CGH for each holder of a national credit card), then the amount of time required to design (and also to fabricate) each CGH should be on the order of 1 s.[2] Computationally intensive design algorithms may also not be appropriate for many optical processors based on spatial light modulators (SLM's); especially, adaptive processors that incorporate new information into newly designed SLM modulations on the fly. It is the later time-critical applications that the algorithms presented in this paper are designed to address.

Methods that we refer to as encoding are especially suitable for fast design of modulations because they calculate the mapping between each desired complex value and each modulator pixel in sequence [see Fig. 1(a)]. Encoding was the principal method of designing CGH's before 1973. With the work of Gallagher and Liu on iterative encoding,[3] there has been a continuing use and refinement of computationally intensive optimization and global search methods to design Fourier transform holograms. For time-critical applications we believe that the advantages of speed and flexibility of encoding are preferable to the performance advantages (especially, diffraction efficiency) of the slower iterative methods.

A second important aspect of encoding onto real-time SLM's is that today's SLM's have far fewer pixels (i.e., space–bandwidth product) than their earlier counterparts, the fixed-pattern CGH pen plots. For the earlier CGH algorithms, it was reasonable to cluster or group pixels together, thereby reducing the space–bandwidth product of the encoded signal by a factor equal to the number of pixels in each group. However, given the low pixel count and the relatively large cost of current SLM's, it is important to utilize as much of the space–bandwidth product of the SLM as possible.

These two considerations on computational speed and bandwidth utilization led to the development of pseudorandom encoding,[4–6] a class of algorithms that encode individual complex values to individual SLM pixels. Since each given value encodes to an individual pixel rather than a group, the space–bandwidth product of the modulation (for periodically sampled SLM's) is identical to the number of pixels in the SLM.

B. Developments in Pseudorandom Encoding Leading toward Ternary Pseudorandom Encoding

The pseudorandom-encoding process represents desired fully complex values (on SLM's that do not produce a complete set of complex values) through the statistical approximation known as the law of large numbers.[7] A unique random distribution of the available pixel modulations is specified for each desired complex value such that the average modulation equals the desired complex value. Under this set of conditions, the resulting Fourier plane intensity pattern will, on average, produce the desired Fourier plane diffraction pattern plus a broadly

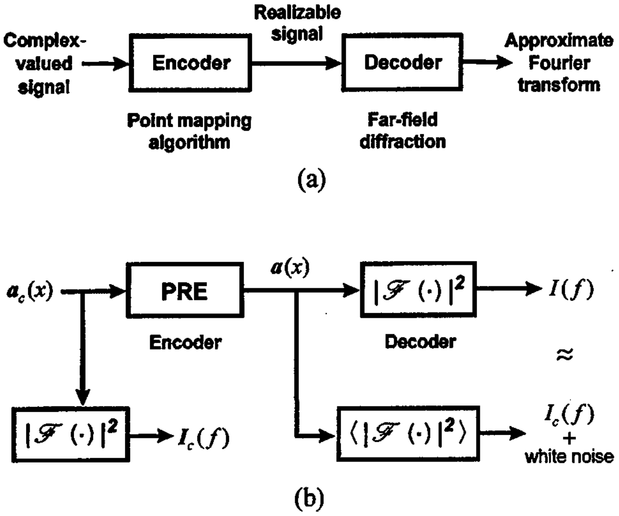


Fig. 1. Systems definition of complex-valued encoding. (a) Systems viewpoint of the Fourier transform computer-generated hologram. A desired complex-valued signal is encoded into a realizable SLM modulation. This signal is decoded through diffraction into a spectrum that approximates the desired complex-valued spectrum. (b) Systems viewpoint specialized for pseudorandom encoding. The desired complex-valued signal $\mathbf{a}_c(x)$ is pseudorandom encoded (PRE) to produce the realizable modulation $\mathbf{a}(x)$. The observed intensity diffraction pattern $I(f)$, which is the squared magnitude of the Fourier transform of $\mathbf{a}(x)$, approximates the desired intensity diffraction $I_c(f)$ in a statistical sense. Specifically, the expected value of $I(f)$, i.e., $\langle I(f)\rangle$, is the desired diffraction pattern $I_c(f)$ on a background of white noise.

spread pedestal that represents the average level of background noise[4] [see Fig. 1(b)]. According to the law of large numbers, the actual diffraction pattern will approximate the average pattern with increasing accuracy as the number of pixels in the SLM is increased.[8] The background noise in the actual pattern is a speckle pattern, which can be either negligible or dominant, depending on the specific complex-valued function that is to be encoded. Some simple metrics calculated from the desired complex function have been described that can be used to provide designers advance knowledge about the quality of each particular encoding.[4,8] Also, in Section 5, a more generally applicable model of signal-to-noise ratio (SNR) is developed.

Until now, pseudorandom-encoding algorithms have always been derived under the assumption that the SLM produces a continuous range of values (e.g., phase-only or coupled amplitude). In one study a continuous (phase-only) modulation has been augmented with a single quantized (zero-amplitude) modulation,[5] but this algorithm does not encode a fully complex range of values if the continuous, unit-amplitude portion of the modulation curve is quantized. However, a key result of this study that is fundamental to the development of pseudorandom encoding on quantized SLM's is the realization that any complex value contained on the line between two complex-valued points can be pseudorandom encoded. By repeatedly using this result for all possible pairs of points on the modulation characteristic, one can identify the encoding range of any given SLM.[6] The complex values identified as encodable by this geometric construction always form a convex set. This analysis procedure was also used to consider the feasibility of pseudorandom encoding on quantized SLM's.[9] The results in Ref. 9 lead to the conclusion that with three properly chosen modulation values a circular region around the origin of the complex plane (i.e., a fully complex range) can be pseudorandom encoded.

This minimal set of restrictions on ternary pseudorandom encoding is of critical importance given the large number of diffractive optics and SLM's that produce only a few quantized levels of modulation.

C. Preliminary Description: Distinctions between Ternary Pseudorandom Encoding and Traditional Computer-Generated Hologram Algorithms

To appreciate better how ternary pseudorandom encoding differs from traditional CGH algorithms, it is worth contrasting it with Burckhardt's method.[10] Burckhardt's method uses a group of three pixels to represent arbitrary complex values. Each pixel is variable in amplitude between zero and unity, and (through delayed sampling) the pixels represent phases of 0°, 120°, and 240°. The addition of these three vector components with various combinations of the three amplitudes permits any value in the hexagonal region shown in Fig. 2(a) to be encoded. The inscribed circle of unity radius is the fully complex set of values that can be encoded by Burckhardt's method.

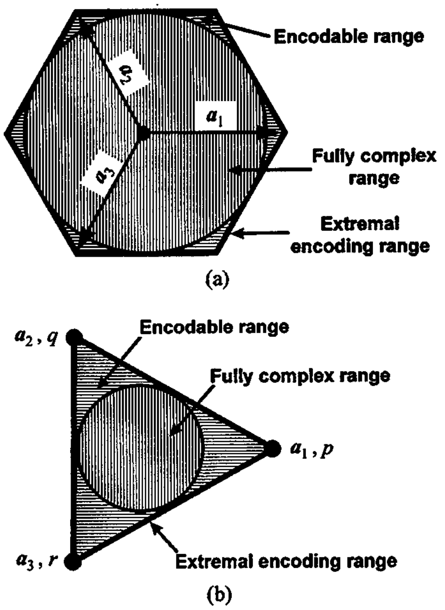


Fig. 2. Distinctions between (a) Burckhardt's ternary encoding method and (b) pseudorandom ternary encoding. In Burckhardt's method the magnitudes of the available complex amplitudes $\mathbf{a}_1$, $\mathbf{a}_2$, and $\mathbf{a}_3$ (their loci indicated by the three connected arrows) can be continuously varied between 0 and 1. In pseudorandom encoding, the available complex amplitudes are constant, but the probabilities p, q, and r of selecting $\mathbf{a}_1$, $\mathbf{a}_2$, and $\mathbf{a}_3$ can be varied continuously between 0 and 1. The constraint that $p + q + r = 1$ leads to pseudorandom encoding having a different encodable/realizable range from that of Burckhardt's method. The fully complex range is the maximum-diameter circular region that surrounds the origin of the complex plane and that does not exceed the extremal encoding range. The fully complex range is drawn for the specific case that the origin is the center of the circular region.

Ternary pseudorandom encoding differs from Burckhardt's method in two key respects: (1) Any one pixel can be set to one of only three complex amplitudes rather than be continuously varied through a range of values, and (2) one pixel, rather than three pixels, is used to represent a desired complex value. Figure 2(b) illustrates the encoding range and the fully complex range (indicated by the inscribed circle) for unit-amplitude modulation at the three values of phase of 0°, 120°, and 240°. The amplitudes that are physically available in Burckhardt's method are effectively represented by the probabilities p, q, and r in ternary pseudorandom encoding. If one of the three possible values of modulation is randomly selected with relative frequencies of occurrence p, q, and r, then the average value of modulation will be the vector sum of the three available modulation values scaled by the respective values of probability. As shown in Ref. 4, this statistical average does effectively represent the pixel modulation for purposes of designing Fourier transform holograms.

The remainder of this paper is organized as follows. Section 2 gives the essential mathematical background on pseudorandom encoding and its properties needed to derive encoding algorithms for quantized SLM's. Section 3 derives the ternary pseudorandom-encoding algorithm. Section 4 evaluates the encoding error for ternary encoding. Sections 3 and 4 also present geometric interpretations of encoding and encoding error. Section 5 describes how ternary encoding can be used to build up encoding algorithms for quantized modulation characteristics and specifically compares the encoding errors for three, four, and five levels of quantization. Section 5 also develops a new model of SNR in terms of measures of the signal to be encoded and the SLM characteristics. Section 6 encodes the same complex function by the various algorithms and compares the resulting simulated and experimentally produced diffraction patterns.

2. MATHEMATICAL BACKGROUND

A. General Description of Pseudorandom Encoding

All pseudorandom-encoding algorithms specify the modulation of any given pixel in terms of a user-specified random variable. The statistical properties of the random variable are selected in such a way that the expected value, or average, of the random modulation is identical to the desired, but unobtainable, fully complex value. The desired complex-valued modulation is written as $\mathbf{a}_c = (a_c, \psi_c)$, and the resulting modulation by the SLM is $\mathbf{a} = (a, \psi)$, where the ordered pairs are the polar representations of the complex quantities. Complex quantities are indicated by bold-face type. The pseudorandom-encoding design statement is, in general, to find a value of the ensemble average

$$\langle \mathbf{a} \rangle = \int \mathbf{a} p(\mathbf{a}) \mathrm{d}\mathbf{a} \qquad (1)$$

of the random variable $\mathbf{a}$ such that $\langle \mathbf{a} \rangle = \mathbf{a}_c$. The statistical properties of $\mathbf{a}$ are determined by its probability-density function (pdf) $p(\mathbf{a})$. The pdf is specified to ensure that the expected value of $\mathbf{a}$ and the desired complex value are identical. After an appropriate density function is determined, the desired complex value $\mathbf{a}_c$ is encoded by drawing a single value of $\mathbf{a}$ from a random distribution having the density function $p(\mathbf{a})$. Since the value of $\mathbf{a}$ is found deterministically by computer, rather than from a random process occurring in nature, the procedure has been named pseudorandom encoding.

This pseudorandom-encoding prescription is applied to each pixel in sequence to encode the desired spatially varying complex modulations $\mathbf{a}_c$. With i as the spatial coordinate, the spatial samples of the desired complex function, the pdf, and the random modulation can be written as $\mathbf{a}_{ci}$, $p_i(\mathbf{a}_i)$, and $\mathbf{a}_i$. This indexing scheme can be conveniently applied to one- or two-dimensional arrays, and it is not restricted to equally spaced samples.

The far-field diffraction pattern of the encoded modulation $\mathbf{a}_i$ approximates the desired diffraction pattern [see Fig. 1(b)]. This can be seen by comparing the intensity of the desired far-field diffraction pattern with the ensemble average diffraction pattern that would result from the encoded modulation. The intensity of the desired diffraction pattern is

$$I_c(f) = \left| \sum_i \mathbf{A}_{ci} \right|^2 = \left| \mathscr{F}\left\{ \sum_i \mathbf{a}_{ci} \right\} \right|^2, \qquad (2)$$

where $\mathscr{F}\{\cdot\}$ is the Fourier transform operator, $\mathbf{A}_{ci}(f)$ is the Fourier transform of the transmittance of the ith pixel located at position i in the modulator plane, and f is the spatial coordinate across the Fourier plane. The expected intensity of the diffraction pattern from the encoded modulation has been derived for the condition that the random variable $\mathbf{a}_i$ for the ith pixel is statistically independent of $\mathbf{a}_j$ for all j not equal to i. Under the pseudorandom design condition $\langle \mathbf{a}_i \rangle = \mathbf{a}_{ci}$, the ensemble average pattern is expressed as[4–6]

$$\langle I(f) \rangle = I_c(f) + \sum_i (\langle |\mathbf{A}_i|^2 \rangle - |\mathbf{A}_{ci}|^2), \qquad (3)$$

where $\mathbf{A}_i(f)$ is the Fourier transform of $\mathbf{a}_i$. The expected intensity consists of two terms. The first term is the desired diffraction pattern from Eq. (2). The second term corresponds to the average level of background (i.e., speckle) noise that is produced as a result of the randomness of the modulation. It is the error signal referred to in Subsection 1.C. For the case of pixels that are modeled as pointlike apertures, the average background noise is of constant intensity for all frequencies f (i.e., it is white).

B. Encoding Error Defined

Equation (3) identifies individually the noise contribution of each pixel. Therefore insight can be gained by evaluating the noise contribution in the modulation plane. Under the assumption that the pixels are infinitesimally wide apertures, the inverse Fourier transform of a single pixel noise term gives the encoding error[6]

$$\epsilon = \langle |\mathbf{a}|^2 \rangle - |\mathbf{a}_c|^2, \qquad (4)$$

where the subscript has been dropped to simplify presentation. (If the pixels are of finite width, then an autocorrelation of the pixel aperture function would also be included in the formula.[4] This term is dropped because it adds no essential insight to the current discussion.)

C. Geometric Interpretation of Binary Pseudorandom Encoding

As stated in Subsection 1.B, the pseudorandom selection between two complex values permits any value on the line segment connecting these two points to be realized on average. This geometric construction was used to determine the encoding range of continuous-range SLM's[6,9] and to identify the pseudorandom-encoding range of quantized SLM's.[9] These results are reviewed here and used to develop insights into pseudorandom encoding with three (or more) quantized levels.

Binary encoding is directly developed by using the pdf for the binary distribution in Eq. (1). The binary pdf is

$$p(\mathbf{a}) = p\,\delta(\mathbf{a} - \mathbf{a}_1) + q\,\delta(\mathbf{a} - \mathbf{a}_2), \tag{5}$$

where $\delta(\cdot)$ is the Dirac delta function, $\mathbf{a}_1$ and $\mathbf{a}_2$ are a pair of complex values from the modulation characteristic, and p and $q = 1 - p$ are the probabilities of selecting $\mathbf{a}_1$ and $\mathbf{a}_2$. Since p is a probability, its value is between 0 and 1. [It will be clear from usage when we are referring to the binary probability p and the density function $p(\mathbf{a})$.] Evaluating Eq. (1) with this pdf gives an expression for the effective complex amplitude of

$$\langle \mathbf{a} \rangle = p\mathbf{a}_1 + (1 - p)\mathbf{a}_2. \tag{6}$$

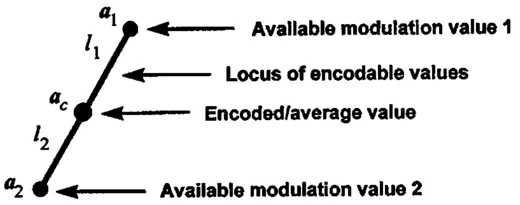
Equation (6) is recognized as the expression for a line as a function of the variable p. For $p = 1$, $\mathbf{a}_1$ is encoded; for $p = 0$, $\mathbf{a}_2$ is encoded; and for values of p between 0 and 1, any value lying on the line segment between $\mathbf{a}_1$ and $\mathbf{a}_2$ can be encoded. This geometric interpretation (see Fig. 3) can be brought out further by considering that the desired complex value $\mathbf{a}_c$ can be expressed in terms of the two complex values $\mathbf{a}_1$ and $\mathbf{a}_2$ as

$$\mathbf{a}_c = \frac{l_2}{l}\,\mathbf{a}_1 + \frac{l_1}{l}\,\mathbf{a}_2, \tag{7}$$

where l_1 is the distance between $\mathbf{a}_c$ and $\mathbf{a}_1$, l_2 is the distance between $\mathbf{a}_c$ and $\mathbf{a}_2$, and $l = l_1 + l_2$. Clearly, the lengths can be chosen so that the desired value $\mathbf{a}_c$ can be realized by the average (or effective) value $\langle \mathbf{a} \rangle$. Evaluation of Eq. (4) by using Eqs. (5) and (7) (and some further algebraic manipulation[9]) shows that the encoding error for binary encoding,

$$\epsilon = l_1 l_2 = pql^2, \tag{8}$$

is simply the products of the distances from $\mathbf{a}_1$ and $\mathbf{a}_2$ to $\mathbf{a}_c$. The maximum encoding error $0.25l^2$ occurs if $\mathbf{a}_c$ is the midpoint of the line segment between $\mathbf{a}_1$ and $\mathbf{a}_2$.



a_1
Available modulation value 1
l_1
Locus of encodable values
a_c
Encoded/average value
l_2
a_2
Available modulation value 2

Fig. 3. Geometry of pseudorandom biamplitude encoding. Any desired value $\mathbf{a}_c$ between the two available modulation values $\mathbf{a}_1$ and $\mathbf{a}_2$ can be encoded by pseudorandom encoding. The product of the lengths of the line segments that connect $\mathbf{a}_1$ and $\mathbf{a}_2$ to $\mathbf{a}_c$ is the encoding error $\epsilon = l_1 l_2$.

Equations (6) and (7) suggest the following encoding formula for binary SLM's: For a given value of p, the desired complex value $\mathbf{a}_c = \langle \mathbf{a} \rangle$ is represented (i.e., encoded) by a single randomly selected value

$$\mathbf{a} = \begin{cases} \mathbf{a}_1 & \text{if } 0 \leqslant s \leqslant p \\ \mathbf{a}_2 & \text{if } p < s \leqslant 1, \end{cases} \tag{9}$$

where s is a uniformly distributed random number between 0 and 1.

D. Using Binary Encoding to Evaluate Encoding Range

The analysis and the geometric interpretation of binary pseudorandom encoding [Eqs. (6)–(8)] provides insight into pseudorandom encoding for various continuous and quantized modulation characteristics.[9] One use of this analysis is in determining those complex values that can be pseudorandom encoded for a particular modulation characteristic. As mentioned in Subsection 1.B, the range is found by combining the ranges encoded by each possible pair of values from the SLM characteristic.[6] Because the binary encoding algorithm has the fewest constraints, the maximum possible range of values (a convex set) is found by this procedure. Other constraints [e.g., using a nonbinary pdf in Eq. (1)] are known to reduce this range.[6] Figure 2(b) shows the convex region (triangular shaded) that is bounded by the three possible binary encodings $\mathbf{a}_1 - \mathbf{a}_2$, $\mathbf{a}_2 - \mathbf{a}_3$, and $\mathbf{a}_3 - \mathbf{a}_1$. This is the range of possible complex values that can be realized with three quantized values of modulation. Also, the circular shaded region represents the range over which fully complex-valued functions can be encoded. Section 3 derives ternary encoding and further shows that there is only one unique solution for encoding a given complex value.

3. TERNARY PSEUDORANDOM ENCODING

The ternary pdf for the three modulation values $\mathbf{a}_1$, $\mathbf{a}_2$, and $\mathbf{a}_3$ is

$$p(\mathbf{a}) = p\,\delta(\mathbf{a} - \mathbf{a}_1) + q\,\delta(\mathbf{a} - \mathbf{a}_2) + r\,\delta(\mathbf{a} - \mathbf{a}_3), \tag{10}$$

where the three probabilities p, q, and r of selecting $\mathbf{a}_1$, $\mathbf{a}_2$, and $\mathbf{a}_3$ satisfy

$$p + q + r = 1. \tag{11}$$

Evaluating Eq. (1) with this pdf gives an expression for the effective complex amplitude of

$$\mathbf{a}_c \equiv \langle \mathbf{a} \rangle = p\mathbf{a}_1 + q\mathbf{a}_2 + r\mathbf{a}_3. \tag{12}$$

There is at most one solution for the values of p, q, and r that encodes $\mathbf{a}_c$. This follows from the fact that there are three linear equations in the three unknowns. The real and imaginary parts of Eq. (12) give two of the equations, and the third expresses that the sum of the three probabilities is unity. These equations written in matrix form are

$$\begin{pmatrix} a_{cr} \\ a_{ci} \\ 1 \end{pmatrix} = \begin{bmatrix} a_{1r} & a_{2r} & a_{3r} \\ a_{1i} & a_{2i} & a_{3i} \\ 1 & 1 & 1 \end{bmatrix} \begin{pmatrix} p \\ q \\ r \end{pmatrix}, \qquad (13)$$

where subscripts r and i indicate the real and imaginary parts of the corresponding complex values $\mathbf{a}_c$, $\mathbf{a}_1$, $\mathbf{a}_2$, and $\mathbf{a}_3$. As long as the matrix is nonsingular, Eq. (13) has a single solution. However, it is possible that the values found for p, q, and r could be less than zero or greater than unity. Since these values are probabilities, such solutions cannot be pseudorandom encoded. We will show that this situation corresponds to the value of $\mathbf{a}_c$ being outside the convex region formed by $\mathbf{a}_1$, $\mathbf{a}_2$, and $\mathbf{a}_3$ [see Fig. 2(b)].

The analysis is performed by using Eq. (11) to eliminate r from Eq. (12), which yields

$$\mathbf{a}_c - \mathbf{a}_3 = p(\mathbf{a}_1 - \mathbf{a}_3) + q(\mathbf{a}_2 - \mathbf{a}_3). \qquad (14)$$

This relationship is illustrated in Fig. 4(a). Using the geometry in Fig. 4(a) and choosing the vector $\mathbf{a}_c - \mathbf{a}_3$ as a coordinate axis, we can write Eq. (14) as

$$\begin{pmatrix} 0 \\ l_{c3} \end{pmatrix} = \begin{bmatrix} l_{13} \sin \theta_{13} & -l_{23} \sin \theta_{23} \\ l_{13} \cos \theta_{13} & -l_{23} \cos \theta_{23} \end{bmatrix} \begin{pmatrix} p \\ q \end{pmatrix}, \qquad (15)$$

where l_{13}, l_{23}, and l_{c3} represent the respective lengths from $\mathbf{a}_1$, $\mathbf{a}_2$, $\mathbf{a}_c$ to $\mathbf{a}_3$ and where θ_{13} and θ_{23} are the angles from $\mathbf{a}_c - \mathbf{a}_3$ to $\mathbf{a}_1 - \mathbf{a}_3$ and $\mathbf{a}_2 - \mathbf{a}_3$, respectively. The solution for Eq. (15) is

$$p = \frac{l_{c3}}{l_{13}} \frac{\sin \theta_{23}}{\sin(\theta_{13} + \theta_{23})}, \qquad q = \frac{l_{c3}}{l_{23}} \frac{\sin \theta_{13}}{\sin(\theta_{13} + \theta_{23})}. \qquad (16)$$

The geometry in Fig. 4(a) limits p and q to be positive, since the angles θ_{13} and θ_{23} are positive and their sum (since they are part of the same triangle) is less than 180°. Equations (16) do admit values of p and q that exceed unity.

We consider some special cases to appreciate better the relationship between values of p, q, and r and values of $\mathbf{a}_c$. First, consider cases where $q = 0$. The construction in Fig. 4(a) indicates that $\mathbf{a}_c$ lies on the line defined by $\mathbf{a}_1$ and $\mathbf{a}_3$. Thus $\theta_{13} = 0$. Solving Eq. (15) for $q = 0$ then gives $p = l_{c3}/l_{13}$, and $r = 1 - (l_{c3}/l_{13})$. This result indicates that ternary encoding reduces to binary encoding [see Eq. (7)] if $q = 0$. Therefore $\mathbf{a}_c$ is contained between $\mathbf{a}_1$ and $\mathbf{a}_3$ as long as p and r are contained between zero and unity. Similarly, if $p = 0$, these equations reduce to binary encoding between $\mathbf{a}_2$ and $\mathbf{a}_3$. If $r = 0$, then binary encoding is performed between the points $\mathbf{a}_1$ and $\mathbf{a}_2$. This situation is illustrated in Fig. 4(b). It is interesting to note that the geometric construction forms two triangles that are identical to the outer triangle $\mathbf{a}_1 - \mathbf{a}_2 - \mathbf{a}_3 - \mathbf{a}_1$, except that they are scaled in size by p and q. Figure 4(c) generalizes this construction for cases where $r \neq 0$. From Eq. (11) it is clear that each side of the outer triangle is divided into lengths that are proportional to p, q, and r. Now three triangles contained inside $\mathbf{a}_1 - \mathbf{a}_2 - \mathbf{a}_3 - \mathbf{a}_1$ are apparent that are identical except for their scaling by p, q and r. This discussion shows that as long as $\mathbf{a}_c$ is contained on the boundary of $\mathbf{a}_1 - \mathbf{a}_2 - \mathbf{a}_3 - \mathbf{a}_1$ or inside the enclosed area, it can be pseudorandomly encoded. Values that are outside correspond to probabilities that are less than zero or in excess of unity, which cannot be realized by this statistical procedure.

Once p and q are found by solving Eq. (13) or by using Eqs. (16), then ternary pseudorandom encoding of the desired complex value $\mathbf{a}_c = \langle \mathbf{a} \rangle$ is accomplished by randomly selecting

$$\mathbf{a} = \begin{cases} \mathbf{a}_1 & \text{if } 0 \leqslant s \leqslant p \\ \mathbf{a}_2 & \text{if } p < s \leqslant p + q, \\ \mathbf{a}_3 & \text{if } p + q < s \leqslant 1 \end{cases} \qquad (17)$$

where s is a uniformly distributed random number between 0 and 1.

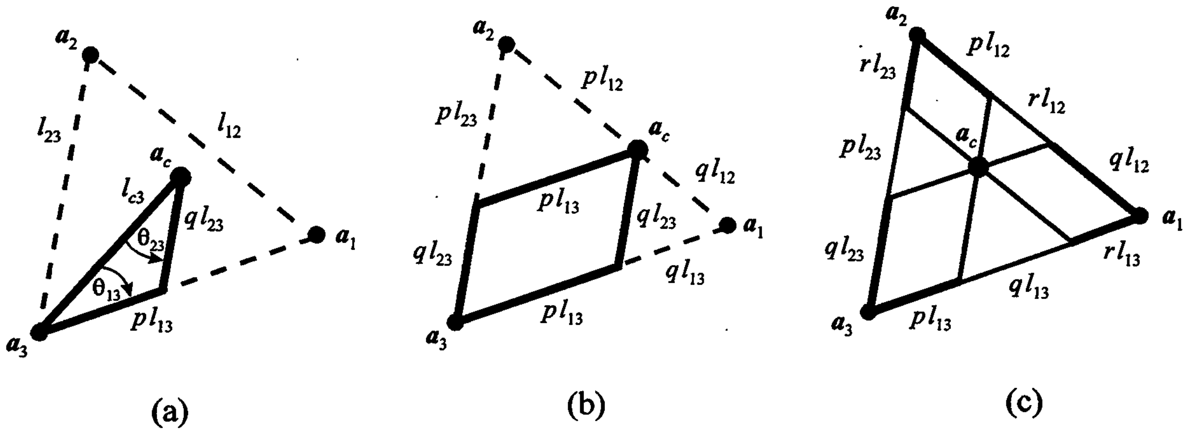


Fig. 4. Geometric relationships for ternary pseudorandom encoding. (a) The three vectors (thick lines) correspond to the three terms in Eq. (14). Each term corresponds to a vector that has $\mathbf{a}_3$ as its origin. (b) Geometry for ternary encoding when the probability $r = 0$. For this condition ternary encoding reduces to biamplitude encoding between $\mathbf{a}_1$ and $\mathbf{a}_2$. This construction also identifies two triangles that are identical except for scaling by p and q. The thick lines indicate the two vectors that add together to produce the desired complex value $\mathbf{a}_c$. (c) Geometry for ternary encoding when the probability r is $0 < r < 1$. This construction shows that there are three triangles that are identical except for scaling by p, q, and r. The thick lines indicate the six line segment lengths that are used in Eq. (22) to calculate the encoding error. The products of the lengths of the three pairs of collinear segments are added together to give the encoding error.

4. ENCODING ERROR AND GEOMETRIC INTERPRETATION

Encoding error provides information on the amount of noise generated by encoding. Since it can be directly calculated from the desired complex value $\mathbf{a}_c$, it can be used to anticipate the quality of the encoding before actually performing the encoding. Therefore pseudorandom encoding has the desirable property that it automatically includes error analysis with the encoding algorithm. In this section the encoding error is evaluated for ternary pseudorandom encoding.

Using Eqs. (10) and (12) in Eq. (4) gives the encoding error for ternary pseudorandom encoding as

$$\epsilon = p|\mathbf{a}_1|^2 + q|\mathbf{a}_2|^2 + r|\mathbf{a}_3|^2 - |p\mathbf{a}_1 + q\mathbf{a}_2 + r\mathbf{a}_3|^2, \tag{18}$$

where the three possible modulation values for a pixel are $\mathbf{a}_1$, $\mathbf{a}_2$, and $\mathbf{a}_3$. Equation (18) can be rewritten as

$$\begin{aligned}\epsilon = (p - p^2){a_1}^2 + (q - q^2){a_2}^2 + (r - r^2){a_3}^2 \\ - 2pqa_1a_2 \cos \phi_{12} - 2pra_1a_3 \cos \phi_{13} \\ - 2qra_2a_3 \cos \phi_{23},\end{aligned} \tag{19}$$

where a_i is the magnitude of $\mathbf{a}_i$ for $i = 1, 2, 3$ and $\phi_{i,j}$ is the angle between $\mathbf{a}_i$ and $\mathbf{a}_j$ for $j = 2, 3$. (Note that in this section the subscripts refer to one of the three possible modulation values for a pixel rather than to spatial position of a pixel.) This result can be dramatically simplified by repeated use of the law of cosines

$$l_{i,j} \equiv |\mathbf{a}_i - \mathbf{a}_j|^2 = {a_i}^2 + {a_j}^2 - 2a_i\, a_j \cos \phi_{ij} \tag{20}$$

and the use of

$$\begin{aligned} p - p^2 &= pq + pr, \\ q - q^2 &= pq + qr, \\ r - r^2 &= pr + qr, \end{aligned} \tag{21}$$

which follows from Eq. (11). Using Eq. (20) three times (for $\{i, j\} = \{1, 2\}, \{1,3\}, \{2, 3\}$) in Eq. (19) together with Eqs. (21) gives the simplified expression for encoding error of

$$\epsilon = pq{l_{12}}^2 + pr{l_{13}}^2 + qr{l_{23}}^2. \tag{22}$$

This result is quite similar to Eq. (8), the encoding error for binary encoding. Since ternary encoding reduces to binary encoding for $r = 0$, then Eq. (22) should also reduce to Eq. (8). Making the identification $l \equiv l_{12}$, we can see that this is indeed the case. Likewise, when $q = 0$ or $p = 0$, the encoding error is identical to the binary-encoding error between $\mathbf{a}_1$ and $\mathbf{a}_3$ or $\mathbf{a}_2$ and $\mathbf{a}_3$, respectively. In general, for p, q, and r not equal to zero, each of the three encoding error terms corresponds to the product of a pair of lengths on the respective line segments $\mathbf{a}_1$–$\mathbf{a}_2$, $\mathbf{a}_1$–$\mathbf{a}_3$, and $\mathbf{a}_2$–$\mathbf{a}_3$. To help visualize the lengths that contribute to encoding error, Fig. 4(c) indicates the lengths as thick lines.

5. DESIGN AND EVALUATION OF SPECIFIC ALGORITHMS

Encoding algorithms for modulation characteristics of any degree of quantization can be built up out of elementary ternary pseudorandom-encoding algorithms. Therefore the analysis presented in Sections 3–5 can be specialized and applied to m-ary quantized modulation characteristics. In this section we define specific pseudorandom-encoding algorithms for three, four, and five levels of quantization that each provide a circular encoding range around the origin of the complex plane. The encoding error for each algorithm is evaluated, and these results are compared with the encoding errors for continuous phase-only and biamplitude phase modulation characteristics. We also use the encoding error together with measures of the desired complex-valued signal to define an estimate of SNR of the resulting diffraction pattern. In Section 6 these specific algorithms are demonstrated for a specified function, and the SNR's of the resulting diffraction patterns are compared with our estimated SNR.

The specific pseudorandom-encoding algorithms evaluated in the remainder of this paper are defined with the help of Table 1. The ternary algorithm is defined to use three modulation values that are equally spaced by $2\pi/3$ rad around the unit circle. The m-ary 1 algorithm uses four modulation values that are equally spaced by $\pi/2$ around the unit circle, and the m-ary 2 algorithm uses the same four modulation values as those for m-ary 1, with the addition of the value of zero. The m-ary 1 algorithm is built up out of two ternary encoding algorithms. The modulation values of $\{1, j, -1\}$ are used for desired complex values that lie in the upper half of the complex plane, and $\{1, -j, -1\}$ is used for encoding complex values that lie in the lower half-plane. The m-ary 2 encoding algorithm is composed of four ternary encoding algorithms (listed in the third column of Table 1). Each of the four ternary algorithms corresponds to encoding desired values in one quadrant of the complex plane.

To compare pseudorandom encoding for discrete modulation characteristics with pseudorandom encoding for continuous characteristics, we also consider encoding algorithms for phase-only[4] and biamplitude phase SLM's.[5] The biamplitude encoding algorithm is identical to that

Table 1. Defining Parameters and Metrics for Various Pseudorandom-Encoding Algorithms

SLM Type	SLM Values	Ternary Groups	γ	ϵ
Ternary	$1, \exp(\pm j2\pi/3)$	$\{1, \exp(\pm j2\pi/3)\}$	$1/2$	$1 - {a_c}^2$
m-ary 1	$\pm1, \pm\mathrm{j}$	$\{1, j, -1\}, \{1, -j, -1\}$	$\sqrt{1/2}$	$1 - {a_c}^2$
m-ary 2	$0, \pm1, \pm j$	$\{0, \pm1, j\}, \{0, 1, \pm j\}$	$\sqrt{1/2}$	$p + q - {a_c}^2$
Phase-only	$\exp(j\psi)$	—	1	$1 - {a_c}^2$
Biamplitude phase	$0, \exp(j\psi)$	—	1	$a_c - {a_c}^2$

reported in Ref. 5. It can be viewed as binary pseudorandom encoding (as described in Section 2) by using the two modulation values of 0 and $\exp(j\psi_c)$, where ψ_c is the phase of the desired complex value $\mathbf{a}_c$ and the binary selection between the two modulation values is controlled by the probability $p = a_c$. The encoding algorithm for phase-only SLM's used in this study is simpler than the algorithm originally reported in Ref. 4. The new algorithm is based on binary encoding between the two values $\pm\exp(j\psi_c)$. In this case the appropriate probability needed to encode the amplitude a_c is $p = (1 + a_c)/2$.

We choose to use this phase-only encoding algorithm over the algorithm in Ref. 4 because it is somewhat simpler to program. However, very similar results are anticipated by either algorithm, since they both produce identical levels of encoding error. This can be appreciated by evaluating the encoding error with the use of Eq. (4). This equation reduces to $\epsilon = 1 - a_c{}^2$ for any phase-only modulation.[4,6] This is true for both continuous and discrete phase-only modulation characteristics [as can also be shown by evaluating Eq. (18)]. Therefore the phase-only, ternary, and m-ary 1 algorithms all produce identical encoding errors when the same amplitude is encoded. The encoding error for biamplitude phase modulation, $\epsilon = a_c - a_c{}^2$, is given in Refs. 5 and 6. This result also follows from Eq. (8) if $l = 1$ and $a_c = p$, which is the case for biamplitude encoding. For the m-ary 2 algorithm, analysis of Eq. (4) or (18) by using any one of the ternary groups (given in the third column of Table 1) gives the encoding error $\epsilon = p + q - a_c{}^2$. The encoding error for each of the five algorithms is summarized in Table 1.

Even though the three algorithms for phase-only SLM's produce identical encoding errors when the same value is encoded, this does not mean that their performance is identical. The reason is that the circular encoding range [see Section 2 and Fig. 2(b)] is less for our discrete modulation characteristics than it is for our continuous characteristics. The scaling of the desired complex-valued function to fit within the maximum circular radius γ of each modulation characteristic can cause significant differences in the amount of encoding error for the various algorithms. For the case of a SLM that produces M uniformly spaced phase-only modulation values around the unit circle, the maximum circular radius can be expressed as

$$\gamma_M = \cos(\pi/M). \tag{23}$$

This result is determined by considering that the fully complex encoding region intersects the chord connecting nearest-neighbor modulation values at the half-angle π/M between them. The values of γ for our specific encoding algorithms are listed in the fourth column of Table 1.

For many pseudorandom algorithms, scaling the complex values to be smaller than the maximum possible radius γ reduces diffraction efficiency and increases SNR.[6] It is also possible to combine pseudorandom encoding with other algorithms, which permits the complex values to be scaled to be larger than the maximum circular radius.[5] This can produce greater diffraction efficiencies and higher SNR's. Considering these additional two possibilities would needlessly complicate this study. In this paper the fully complex values are scaled so that the maximum amplitude for a given encoding algorithm is its maximum circular radius γ.

With this definition of the maximum circular encoding range, it is now possible to make a comparative analysis of the performance of each algorithm. Since the desired fully complex function $\mathbf{a}_c(x)$ is normalized so that its maximum amplitude is γ, it is appropriate to compare the amplitudes a_{c1}/γ_1 and a_{c2}/γ_2, where the subscripts 1 and 2 indicate the values of amplitude and maximum amplitude for two different algorithms. Since encoding error is proportional to intensity rather than amplitude, it is appropriate to compare the normalized encoding errors $\epsilon_1/\gamma_1{}^2$ and $\epsilon_2/\gamma_2{}^2$. Therefore, rather than describing the absolute encoding error, this normalization presents the error-to-signal ratio, or relative error, for the same value encoded by two different algorithms. Relative error is more informative of the fidelity of the diffraction patterns resulting from pseudorandom encoding than is absolute encoding error.

The encoding errors for the Table 1 algorithms are presented in this way in Fig. 5. Phase-only encoding always produces larger encoding errors than biamplitude phase encoding, as reported in Ref. 5. The m-ary 2 algorithm produces a range of relative errors that are contained between the two m-ary 2 curves in Fig. 5. The lower curve corresponds to the case where $q = 0$, and the upper curve corresponds to the case where $p = q$. Figure 5 shows that the m-ary 2 algorithm always produces more relative error than does biamplitude phase encoding but frequently produces less relative error than does phase-only encoding. If the desired complex-valued function $\mathbf{a}_c(x)$ has many more amplitudes that are less than 1/2, then the total relative error produced by m-ary 2 encoding can be substantially less than that for phase-only encoding. Alternatively, if most of the desired complex values are well above 1/2, then m-ary 2 encoding produces a total relative error that is much greater than that for phase-only encoding.

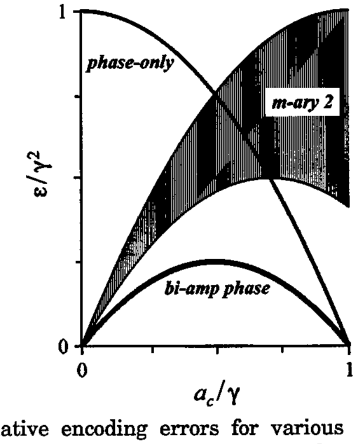


Fig. 5. Relative encoding errors for various pseudorandom-encoding algorithms. The desired magnitude $\mathbf{a}_c$ is normalized by γ, the maximum radius for the fully complex encoding range, and the relative encoding error is $\epsilon_{\rm rel} = \epsilon/\gamma^2$. The striped region gives all possible encoding errors for the m-ary 2 algorithm. The phase-only curve also represents the relative encoding error for pseudorandom encoding with M phase-only values that are uniformly spaced in angle. In this case the plotted phase-only curve is offset by $\tan^2(\pi/M)$ [see Eq. (24)].

The relative error curves for ternary and m-ary 1 encoding are identical to the relative error curve for phase-only encoding, except that they are offset by amounts that depend on the degree of quantization. This can be seen by considering that for quantized phase-only modulation characteristics for which there are M evenly spaced modulation values on the unit circle, the relative error can be written with the help of Eq. (23) as

$$\epsilon_{\mathrm{rel}} \equiv \epsilon_M/\gamma_M{}^2 = 1 + \tan^2(\pi/M) - (a_c/\gamma_M)^2. \quad (24)$$

For the maximum amplitude of $a_c = \gamma_M$, Eq. (24) gives the minimum relative error $\epsilon_{\mathrm{rel}} = \tan^2(\pi/M)$, and for the minimum amplitude of $a_c = 0$, the relative error is $\epsilon_{\mathrm{rel}} = 1 + \tan^2(\pi/M)$. Between these limiting points the curve has the identical quadratic dependence as that of the relative error curve for the continuous phase-only modulation characteristic. Therefore the quantized characteristics produce additional relative error by the amount $\tan^2(\pi/M)$. For the ternary encoding this offset is 3, or in other words, the relative error for ternary encoding is always larger by a value of 3 than that for phase-only encoding. For m-ary 1 encoding the offset is 1. With increasingly fine quantization the relative error curve approaches the phase-only curve. For instance, the offset would be 0.17 for eight levels and 0.04 for 16 levels of quantization.

Despite the significant amount of relative error produced by ternary pseudorandom encoding, it is possible to use this noisiest of pseudorandom algorithms to encode many desired complex functions with good fidelity. The key factor is that the spatial extent (i.e., the bandwidth B) of the desired diffraction pattern is small enough that the signal is sufficiently greater than the background noise that is due to the sum of the encoding errors from each pixel.

A simple analysis is presented to make this relationship more apparent. Consider that a particular desired function is pseudorandom encoded for an N-pixel SLM. The average encoding error per pixel is ϵ_a, and the average intensity transmittance that is encoded is $a_{ca}{}^2$. The total energy in the encoding error is then $N\epsilon_a$, and the total energy in the encoded signal is $Na_{ca}{}^2$. The encoding error in the diffraction plane transforms into a white spectrum over a bandwidth of N, whereas the desired signal has a designed bandwidth of B. Therefore the desired diffraction pattern will have a directionality gain of N/B over the spectrum of the encoding error. The SNR can then be written for this approximate analysis as

$$\mathrm{SNR} = \frac{N}{B}\frac{a_{ca}{}^2}{\epsilon_a}. \quad (25)$$

To appreciate this analysis better, consider the following numerical example. For a 128 × 128-pixel phase-only SLM ($N = 16{,}384$), an average encoding error $\epsilon_a = 0.9$, a root-mean-square amplitude transmittance $a_{ca}{}^2 = 1 - \epsilon_a = 0.32$ (from Table 1), and a desired signal-to-noise ratio SNR $\geqslant 100$, Eq. (25) gives the result that the bandwidth of the desired signal needs to be $B \leqslant 18.2$. This analysis shows that even for pseudorandom algorithms that produce the greatest encoding errors, there are many diffraction patterns that can be successfully encoded as long as the SNR is acceptably large and the signal bandwidth is correspondingly low. The trends predicted by Eq. (25) are evident in the computer simulations and the experimental demonstrations that are presented in Section 6.

Although fidelity of diffraction patterns is the focus of this investigation, diffraction efficiency is probably the most widely discussed metric. For this reason a few basic relationships are reviewed that relate diffraction efficiency to SNR. In pseudorandom encoding, the diffraction efficiency has been shown to be identical to the average intensity of the fully complex function that is to be encoded.[8] This is written as

$$\eta = \frac{1}{N}\sum_{i=1}^{N} a_{ci}{}^2 \equiv a_{ca}{}^2. \quad (26)$$

Since, for a particular encoding method, the a_{ci} are scaled so that the maximum amplitude equals the circular encoding radius γ, it is clear from Eq. (26) that the diffraction efficiency is proportional to γ^2.[6] Consequently, the efficiencies of the five algorithms in Table 1 vary by a factor of 4.

For arbitrary modulation characteristics the average intensity transmittance in Eq. (26) can be replaced by the diffraction efficiency η, and for the specific case of phase-only modulation Eq. (26) reduces to

$$\mathrm{SNR} = \frac{N}{B}\frac{\eta}{1-\eta}. \quad (27)$$

However, for the two modulation characteristics in Table 1 that also have a zero amplitude the denominator term ϵ_a is less than $1 - \eta$, which increases the SNR over that possible for the corresponding discrete or continuous phase-only characteristic.

6. DEMONSTRATIONS OF TERNARY ENCODING

A. Specification of the Desired Function To Be Encoded

In this section the same desired function $\mathbf{a}_{ci}$ is encoded by various algorithms, and the resulting simulated and (for some cases) experimentally measured diffraction patterns are presented. The desired function is a 128 × 128 array of complex values that has a Fourier transform that produces a 7 × 7 array of uniform-intensity diffraction-limited spots. To better compare pseudorandom encoding with known art, we relate our results to previously published designs. Krackhardt *et al.* have reported the highest possible diffraction efficiency designs for continuous phase-only SLM's.[11] Our complex values are derived from their design for a 1 × 7 spot array. Their Table III specifies seven phases θ_k associated with seven equally spaced spots. These phases are used to specify a desired fully complex function of the form

$$\mathbf{a}_c(x, y) = \sum_{k=2}^{8} \exp(j\theta_k)\exp(j2\pi kx) \times \sum_{l=2}^{8} \exp(j\theta_l)\exp(j2\pi ly) \quad (28)$$

that is rectangularly separable. This periodic function is sampled to produce a 32 × 32-unit cell of complex values and a 4 × 4 array of cells that form the 128 × 128 desired complex values.

B. Definition of the Encoding Algorithms Used

These values are then encoded by each of the five algorithms in Table 1. (Part of the encoding includes scaling the desired complex values so that the largest amplitude of the complex values is equal to the appropriate value of γ given in Table 1.) The resulting diffraction patterns are compared and evaluated.

Equation (28) can also be interpreted as a desired fully complex function for the study of Krackhardt *et al.* The phases found through their global optimization of the diffraction pattern specify a desired function that is then encoded onto continuous phase-only SLM's by transforming the desired fully complex function into the phase-only function or kinoform[12]

$$\mathbf{a}(x, y) = \exp\{j \arg[\mathbf{a}_c(x, y)]\}. \tag{29}$$

This encoding is performed not only so that the desired function can be implemented by a phase-only modulator but also to maximize diffraction efficiency.[13] The encoding indicated in Eq. (29) is also applied to the 128 × 128 desired complex values, and the resulting diffraction pattern is compared with those of the five pseudorandom algorithms. We will refer to this algorithm as nonrandom phase-only encoding to help distinguish it from pseudorandom phase-only encoding.

Since the algorithms in which we are most interested are for quantized SLM's, we also quantize the phase to three and four values of phase (uniformly spaced around the unit circle) and evaluate the diffraction patterns for these modified encodings. We refer to these algorithms as nonrandom ternary and nonrandom m-ary 1, respectively. It should be noted in Tables I and II of Ref. 11 that different values of the spot phases θ_k would lead to maximum diffraction efficiency. However, this would require a new optimization to find the phases for each modulator characteristic. Instead, in keeping with the spirit of encoding the same complex function, we have chosen to compare quantized pseudorandom encoding with quantized versions of the maximum-efficiency, rectangularly separable design.

It is also possible to specify a nonrandom version of the pseudorandom m-ary 2 algorithm. In this case a zero-value modulation is selected if the desired complex value is closer to zero than to the four other phase-only modulation values. To compare the pseudorandom and nonrandom m-ary 2 algorithms fairly, the desired complex values $\mathbf{a}_{ci}$ are similarly scaled, so that the maximum amplitude encoded is $\gamma = \sqrt{1/2}$. However, Juday has shown that the quality of an encoding depends on the value of γ.[5,14] For this reason we also perform an iterative search to find the value of γ that optimizes the performance measures that are of most interest to us. The optimum value found in our simulations is $\gamma = 1.3$, which optimizes signal-to-peak-noise ratio (SPR) and uniformity. Subsection 6.C defines these two and the other metrics of interest.

C. Simulation Procedures and Definition of the Performance Metrics

Two metrics, diffraction efficiency η and signal-to-noise ratio SNR, are directly calculated from the desired complex values $\mathbf{a}_{ci}$ for each of the five pseudorandom-encoding algorithms. After $\mathbf{a}_{ci}$ is scaled by the appropriate value γ in Table 1, η is calculated by using Eq. (26) and SNR is calculated by using Eq. (25). In these calculations we use $N = 128^2$ for the number of SLM pixels, and, considering a diffraction-limited spot to have a space–bandwidth product of 1, we use $B = 7^2$ for the space–bandwidth product of the desired signal. These metrics that are based on theory are listed in parentheses in Table 2 beside the values of SNR and η, which are calculated directly from the simulated diffraction patterns.

The far-field diffracted intensity patterns are simulated by fast-Fourier-transforming the encoded values $\mathbf{a}_i$ and then squaring the magnitude for each of the pseudorandom and nonrandom encodings. For all metrics except diffraction efficiency, the 128 × 128 array is placed in a 512 × 512 array of zeros that is fast Fourier transformed. The zero padding is used to resolve the features of the diffraction pattern more finely and to produce more realistic gray-scale images. For diffraction efficiency the 128 × 128 array is fast Fourier transformed directly. For phase-only modulation (for either pseudorandom or nonrandom algorithms), the efficiency is simply the sum of the intensities of the 49 spots divided by the sum of all intensities in the 128 × 128 diffraction pattern. For the other modulation characteristics that contain a zero value, the energy absorption in the modulator plane also needs to be accounted for.[5] Therefore the ratio of desired energy to total energy in the diffraction pattern is multiplied by the ratio of unit-amplitude pixels (i.e., "on" pixels) to the total number of pixels (i.e., number of on plus off pixels). The SNR is the ratio of the average intensity of the peak values of each of the 49 spots to the average intensity of the 512 × 512 pattern, excluding the square

Table 2. Performance Measures of the Pseudorandomly Encoded Desired Function

Pseudorandom	η (%)[a]	SNR[b]	SPR	n-unif (%)
Biamplitude phase	44 (44)	685 (710)	38	4
m-ary 2	22 (22)	197 (203)	15	10
Phase-only	43 (44)	254 (262)	17	10
m-ary 1	22 (22)	93 (94)	8	11
Ternary	11 (11)	40 (41)	3	20

[a] Numbers in parentheses are calculated from Eq. (26).
[b] Numbers in parentheses are calculated from Eq. (25).

Table 3. Performance Measures of the Nonrandomly Encoded Desired Function

Nonrandom	η (%)	SNR	SPR	n-unif (%)
m-ary 2 ($\gamma = \sqrt{1/2}$)	7	175	1	60
m-ary 2 ($\gamma = 1.3$)	68	961	5	11
Phase-only	92	1076	77	18
m-ary 1	75	1189	5	18
Ternary	63	376	1	18

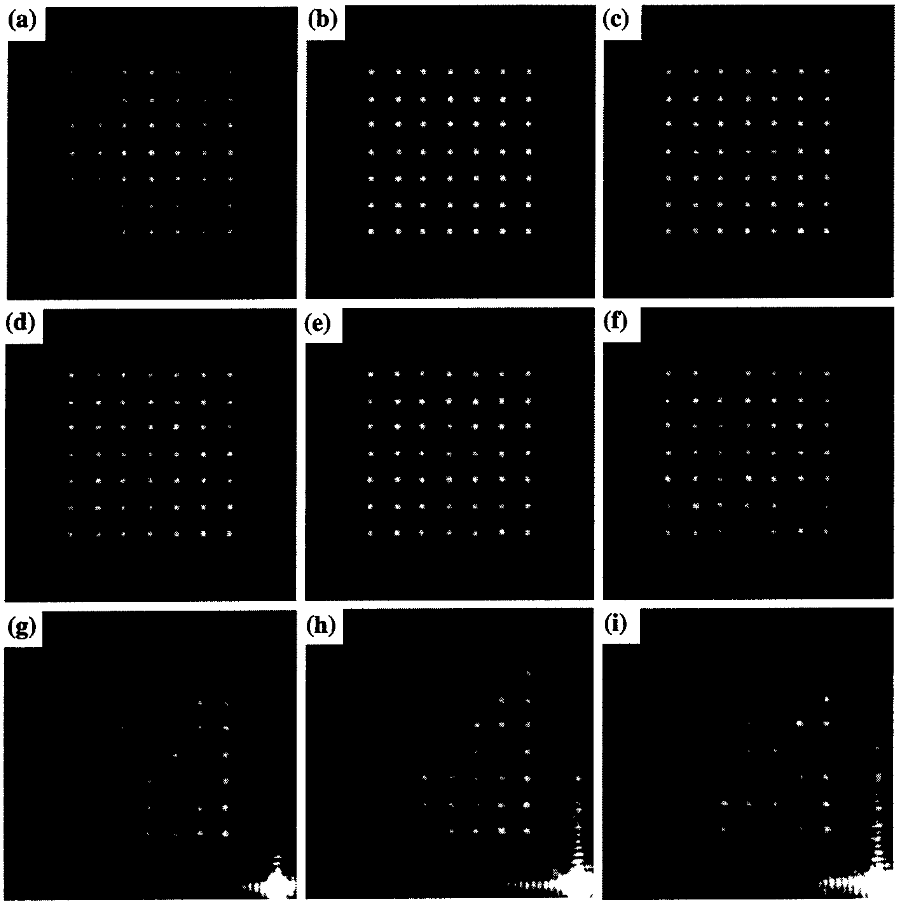


Fig. 6. [(a)–(f)] Simulated and [(g)–(i)] experimental gray-scale images of the diffraction pattern intensity resulting from various encoding algorithms. All encodings are pseudorandom except (a), which is nonrandom phase-only. The simulated pseudorandom encodings are (b) biamplitude phase, (c) m-ary 2, (d) phase-only, (e) m-ary 1, and (f) ternary. The experimental pseudorandom encodings are (g) phase-only, (h) m-ary 1, and (i) ternary.

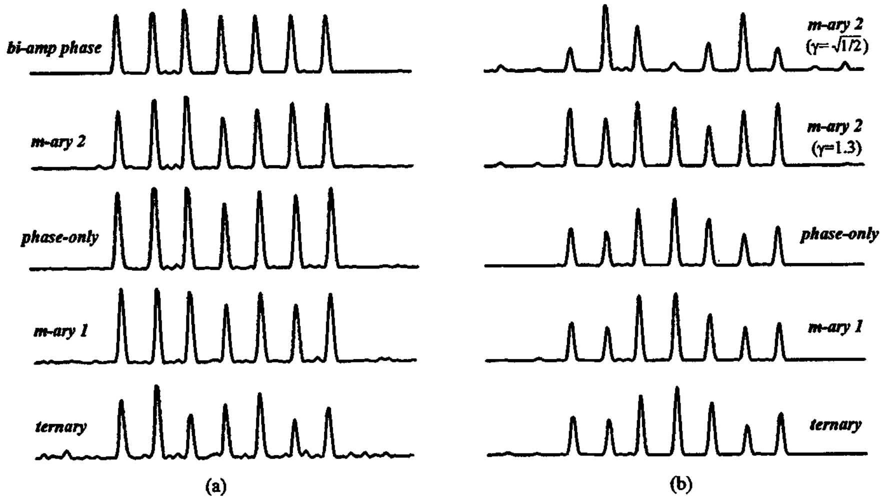


Fig. 7. Simulated cross sections of the diffraction pattern intensity resulting from various encoding algorithms: (a) pseudorandom encodings, (b) nonrandom encodings. Each cross section is along a diagonal that contains the (7,7) order (leftmost spot), the (1,1) order (rightmost spot), and the optical axis (the rightmost side of the curve). For all the pseudorandom curves in (a) and the nonrandom phase-only curve in (b), the cross sections are the intensity values from the diagonal (from upper left to lower right) of the corresponding gray-scale image in Fig. 6. The four other nonrandom encodings in (b) are plotted over an identical range.

region that contains the 7 × 7 spot array. The SPR is the ratio of the average peak intensity of the spots to the maximum noise peak outside the square region containing the spots. Nonuniformity of the peaks (abbreviated in Tables 2 and 3 as n-unif) is well characterized by the standard deviation of the peak intensities of the 49 spots divided by the average spot intensity. We present the nonuniformity metric this way instead of as the maximum peak-to-peak fluctuation of the spots because this metric is less susceptible to the random variations that occur for each new set of random numbers used in the encoding algorithm. We did observe that the peak-to-peak fluctuations are 1.8–2.5 times greater than the values of nonuniformity reported in Table 2 and below in Table 3.

D. Simulation Results

Figures 6(a)–6(f) show the same portion of the 512 × 512 diffraction patterns for nonrandom phase-only encoding [Fig. 6(a)] and for each of the five pseudorandom encodings [Figs. 6(b)–6(f)]. Figure 7(a) shows a cross section of intensity pattern [along a diagonal slice containing the optical axis and the (1, 1) and (7, 7) spots] for the five pseudorandom-encoded patterns. Figure 7(b) shows the corresponding cross sections for the nonrandom encodings. The diagonal cross sections tend to accentuate intensity fluctuations between the spots and to exclude most of the sidelobes between the spots that are due to the sinc-squared nature of the diffraction pattern of each spot. The three patterns of Figs. 6(d)–6(f) show increasing levels of background speckle noise and nonuniformity of the spots. The pseudorandom biamplitude phase encoding [Fig. 6(b)] appears to be free of speckle noise, and the quantized biamplitude encoding [i.e., m-ary 2, Fig. 6(c)] appears to have a noise level that is indistinguishable from that of the phase-only encoding [Fig. 6(d)]. Although most of the cross sections in Fig. 7(a) are reasonably uniform, it is clear that the ternary encoding is the least uniform of the five pseudorandom encodings. From the metrics in Table 2, it is possible to make the following observations about pseudorandom encoding:

1. Though not unexpected, Table 2 demonstrates that as the quantization becomes coarser, the performance decreases. This observation also applies to comparisons between phase-only and biamplitude phase encoding.
2. The small observed differences between the gray-scale images for the m-ary 2 [Fig. 6(c)] and phase-only [Fig. 6(d)] encodings are borne out by the relative differences between their metrics in Table 2. The only significant difference is in diffraction efficiency, which is not reflected in the gray-scale images.
3. The theoretical diffraction efficiency of Eq. (26) and the SNR of Eq. (25) are in close agreement with the simulated results.
4. The values of SNR are 12–18 times larger than those of SPR. The much lower SPR is probably due in large part to the background speckle noise being exponentially distributed in intensity.[15] The exponential pdf decreases to zero very slowly with increasing values of intensity. There are also on the order of $N = 16{,}384$ (the number of pixels and also the space–bandwidth product of the speckle noise pattern) independent noise samples in the diffraction pattern. In 16,384 independent Bernoulli trials,[16] there is a probability of approximately 50% that the maximum value is 10× greater than the average value of an exponential distribution, and there is a 10% probability that the maximum value is 12× greater than the average. This leads to the possibility of the maximum-valued noise peak used to calculate SPR being substantially larger than the average value of noise intensity used to calculate SNR. However, since our maximum noise peaks always tend to be somewhat larger than the 50th percentile, there appear to be other contributions to the background that we have not been able to account for.

The model is presented mainly to provide insight into the noise properties of pseudorandom encoding. These properties are noticeably different from those for nonrandom encoding, as we show presently.

The nonrandom-encoding algorithms can all be viewed as a point-by-point nonlinear transformation of the desired complex values. This type of nonlinearity usually produces mixing products that appear as unwanted sum and difference frequencies in the diffraction pattern and as interference in the signal strengths of the desired frequencies.[2] The mixing products are not very evident in the gray-scale image for nonrandom phase-only encoding [Fig. 6(a)]. The interference, which perturbs the uniformity of the spot array, is evident in the intensity curves for phase-only as well as for m-ary 1 and ternary encodings in Fig. 7(b). The m-ary 2 ($\gamma = 1.3$) encoding is the most uniform of the nonrandom encodings shown. Some mixing products at frequencies other than for the desired 7 × 7 spot positions are also apparent in all the intensity curves except for the nonrandom phase-only encoding. However, the mixing products are significantly larger away from the desired spot array, as shown in Fig. 8(a) for the nonrandom ternary encoding. The spot array is designed to lie to the upper left of the optical axis, which gives rise to the strong unwanted harmonics in the lower right corner of the gray-scale image. All the other nonrandom encodings are similar in that the noise is primarily distributed in this same spatial pattern (though with differing intensity levels).

Comparing the simulated nonrandom encodings with pseudorandom encodings, we observe that the pseudorandom encodings all have a speckle/noise pattern of the same average intensity and visual texture over the entire simulated diffraction pattern, as is shown in the close-up views of Figs. 6(b)–6(f). The intensity cross sections for the nonrandom encodings in Fig. 7(b) generally appear less uniform than those for the corresponding pseudorandom curves in Fig. 7(a). The background noise is much more evident on the pseudorandom ternary encoding [Fig. 7(a)] than it is on the nonrandom ternary encoding [Fig. 7(b)]; however, the peak noise of the nonrandom encoding [shown over a wider spatial extent in the cross section in Fig. 8(a)] is significantly larger than the background noise for pseudorandom encoding.

Table 3 provides more detailed information for comparing the individual nonrandom-encoding algorithms with each other and with the results in Table 2 for pseudorandom encoding. The nonrandom diffraction efficiencies are generally much higher than those for pseudorandom encoding. However, the m-ary 2 ($\gamma = \sqrt{1/2}$) encoding has an extremely low diffraction efficiency. This is a direct result of the small value of the maximum complex radius γ, which leads to most of the desired complex values being closer to zero than to a unity magnitude point. This led to our use of the m-ary 2 ($\gamma = 1.3$) encoding, which produces much more uniform spot arrays than does the nonrandom phase-only encoding.

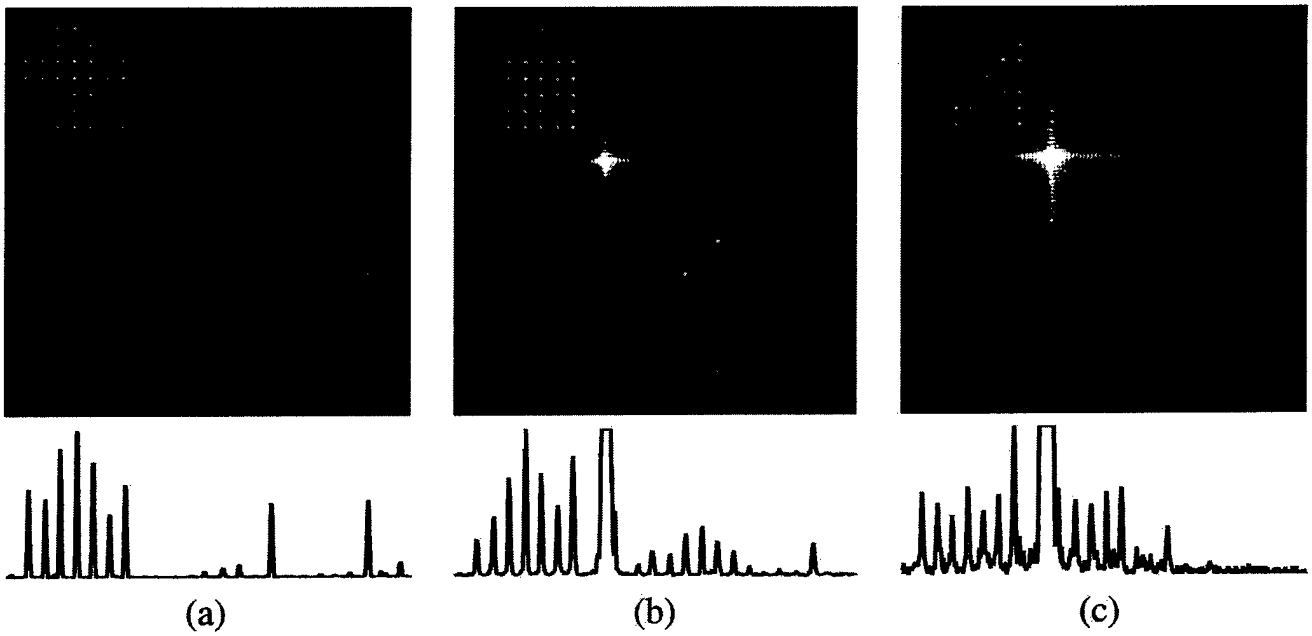


Fig. 8. Delineation of nonlinear effects on encoding: (a) simulated and (b) experimental diffraction pattern intensity for nonrandom ternary encoding, (c) experimental diffraction pattern for pseudorandom ternary encoding. These patterns show a larger view of the diffraction pattern than those in Figs. 6 and 7. Each intensity cross section is along the diagonal of the corresponding gray-scale image. In (a) and (b) the nonrandom ternary encoding produces mixing products, as evident in the lower left corner of each gray-scale image. Although speckle noise is evident in this same region for pseudorandom ternary encoding [(c)], it is much lower in intensity than the mixing products for (b). The saturated spot (centered on the optical axis) in (b) and (c) is primarily a result of the SLM cover glass not being antireflection coated. The most severe effect of the SLM's limited resolution is the appearance, to the lower left of the optical axis, of a duplicate 7 × 7 spot array in (b) and (c).

For completeness Table 3 reports SNR. This number has little practical use, since much of the noise energy appears in a relatively few noise spikes. This leads to the ratio of SNR to SPR being greater than 100 for four of the five nonrandom algorithms, with the one exception being the maximum-efficiency, phase-only encoding.

The nonrandom phase-only encoding by far produced the highest diffraction efficiency and SPR of any encoding algorithm. In fact, so much energy is in the desired spots that there is little energy left to contribute to noise. However, the spot array is less uniform than for the pseudorandom phase-only encoding. It should be noted that Krackhardt *et al.* reported that by reoptimizing the design, it is possible to lower the diffraction efficiency somewhat, which results in a more uniform array of spots.[11] Nonetheless, pseudorandom phase-only encoding produces reasonably uniform spot arrays and large values of SPR, and it does this without the added computational effort of reoptimizing. This becomes clearer when it is seen that the SPR of the quantized modulation characteristics is much lower for nonrandom encoding (Table 3) than it is for pseudorandom encoding (Table 2). This can be interpreted that pseudorandom encoding is less severely affected (i.e., it is more robust) than nonrandom encoding by quantization effects. The results in Krackhardt *et al.* for binary modulation suggest that reoptimization would be needed for each new type of quantization to minimize these effects. Even with its lower diffraction efficiency, the pseudorandom encoding appears to produce a more faithful reconstruction, with less computational effort, than do the nonrandom encodings. Furthermore, based on the studies in Ref. 5 on biamplitude encoding, it appears likely that there is a way to blend random and nonrandom algorithms for discrete-value modulation such that by adjustment of the value of γ (similar to our optimization of nonrandom m-ary 2 encoding) the uniformity and the SPR are improved over those possible with either random or nonrandom encoding individually. This iteration increases computation, but since only one parameter is adjusted, the computation time should be significantly less than that for the global optimization approaches used in, e.g., Refs. 3 and 11.

E. Procedures Used for the Experiments

We also have attempted to modulate a Hughes birefringent liquid-crystal light valve (set up in a phase-only mode) with the encoded phase values. Cohn *et al.* previously reported experiments on phase-only encoding for this light valve.[15] The current setup differs from the previous setup in that (1) a 488-nm laser is now used in place of a 633-nm laser, (2) the pixel array is now a 128 × 128 array instead of a 100 × 100 array [used for Fig. 3(c) in Ref. 15], and (3) the video signal that drives the write light monitor (an Electrohome EDP58XL monochrome monitor with a Hughes high-brightness red tube) is now derived from a Coreco video display card ($S3$ chip set) set to a resolution of 800 × 600 noninterlaced pixels. Previously, a National Television System Committee (NTSC) signal was the video source. As in Ref. 15, a SLM pixel corresponds to three video lines or 3 × 3 pixels from the video display card. We have characterized the transfer function from gray-scale values in digital memory to the phase modulation of the light valve. With the light valve drive voltage set to 27 V *p.-p.* (2 kHz) and with proper adjustment of the brightness and the contrast of the monitor, we have realized a nearly linear transfer function in which a gray-scale value of 80 corresponds to zero phase shift and a gray-scale value of 255 corresponds to a phase shift of 2π. The monitor magnification is set to minimum in the horizontal direction. The monitor is reimaged with a 1.9× reduction onto the write side of the light valve. The resulting image is 21 mm × 21 mm, which we have determined to be the maximum input aperture size that allows us to produce diffraction-limited optical Fourier transforms. This area is illuminated on the read side of the light valve with the light polarized along the extraordinary axis of the liquid crystal. The beam converges to a focal point approximately 2 m from the light valve. A 2033 × 2044-pixel cooled CCD camera placed at the focal point records the resulting diffraction patterns.

F. Experimental Results

Figures 6(g), 6(h), and 6(i) show the diffraction patterns for pseudorandom phase-only, *m*-ary 1, and ternary encoding, respectively. The images differ from the simulated patterns in Figs. 6(d), 6(e), and 6(f), respectively, in that there is a bright spot centered on the optical axis (lower right corner of each image) and that the intensity rolls off/decreases with distance from the optical axis. The bright spot is due primarily to the cover glass of the light valve not being antireflection coated. The roll-off is due to the limited spatial frequency response of the phase of the SLM. The filtering of the phase also produces nonlinear mixing products that contribute energy to the on-axis spot and to an unwanted mirror image of the desired 7 × 7 spot to the lower left of the optical axis in Figs. 8(b) and 8(c). By comparison of Figs. 8(a)–8(c), it can be seen that nonrandom encoding itself produces strong nonlinear mixing products at a few frequencies [at the lower right of Figs. 8(a) and 8(b)]. Instead of producing mixing orders, pseudorandom encoding [Fig. 8(c)] produces an average low level of speckle noise over the entire spatial extent of the diffraction pattern. For either nonrandom- or pseudorandom-encoding experiments, additional nonlinear terms (the mirror images) are present in Figs. 8(b) and 8(c) as a result of the limited phase resolution of the SLM. This distortion, which is due to the limitations of the SLM rather than the encoding method, makes it difficult to make meaningful comparisons between simulation and experiment. Nonetheless, qualitative agreement between Figs. 6(c) and 6(g), 6(d) and 6(h), and 6(e) and 6(i) is seen in that the noise level increases as the quantization becomes coarser.

Potentially much closer agreement between theory and experiment is anticipated by using electrically addressed SLM's. Most of the current devices have individually defined electrodes for each pixel, which would minimize resolution loss that is due to electrostatic fringing fields across the liquid-crystal layer. These SLM's are available only as research-grade or custom devices. In our case this has led to significant delays in obtaining a fully functional device. We have previously observed very close agreement between simulation and theory for pseudorandom phase-only encoded designs that were implemented as diffractive optical elements.[17,18] These results assure us that given an adequately ideal phase-only (or coupled amplitude–phase) SLM, it would be possible to encode fully complex functions, even if the SLM is capable of producing only a few discrete levels of modulation.

7. SUMMARY

In this paper we have derived a statistically based algorithm that with as few as three discrete modulator values encodes a desired complex value to a single pixel in an average sense. This pseudorandom ternary algorithm can be applied directly to SLM's that produce only three values. For SLM's that produce several discrete values, multiple groups of three values can be used to subdivide the complex plane into smaller areas that are ternary encoded, which consequently produces smaller amounts of encoding error. The effect of quantization on pseudorandom encoding is well characterized in a simple model of SNR that depends on only four parameters: two that depend on the signal to be encoded (signal bandwidth and signal diffraction efficiency), one that depends on the modulator characteristic (number of SLM pixels), and one (average encoding error per pixel) that depends on both the signal to be encoded and the modulator characteristic curve. We demonstrated in our simulations that this metric accurately describes the SNR of spot array generators.

To better appreciate the performance of pseudorandom-encoding algorithms, we have compared these algorithms with currently used algorithms in which a desired fully complex function is mapped into modulation values in a systematic and nonrandom way. This function (selected by a global optimization procedure), when mapped to a continuous, phase-only modulation characteristic, produces a diffraction pattern that has the highest diffraction efficiency and the highest SPR of all encoding algorithms studied herein. However, it is less uniform than four of the five pseudorandom algorithms. Furthermore, the SPR of the nonrandom algorithms becomes much worse than that of the pseudorandom algorithms that use similarly quantized modulation characteristics.

Even though the pseudorandom algorithms are less diffraction efficient than nonrandom algorithms, these results nonetheless indicate that pseudorandom algorithms offer significant advantages in terms of fidelity of the diffraction pattern. The advantages of complex-valued encoding techniques, despite their lower diffraction efficiencies, are further amplified by Kettunen *et al.*[19] These advantages, coupled with the low computational overhead of the encoding algorithm and its ability to place a signal anywhere in the available space–bandwidth product of the SLM, make it especially useful for today's low-pixel-count SLM's. We have also observed that the performance advantages of pseudorandom encoding over nonrandom encoding are even more pronounced when it is not possible to maximize the diffraction efficiency of the desired fully complex function, such as in many real-time and time-critical applications.

ACKNOWLEDGMENTS

This study was supported by NASA cooperative agreement NCC5-222 and Office of Naval Research grant N00014-96-1-1296.

Send all correspondence to Robert W. Cohn, The ElectroOptics Research Institute, Room 442, Academic Building, University of Louisville, Louisville, Kentucky 40292; tel, 502-852-7077; fax, 502-852-1577; e-mail, rwcohn01@ulkyvm.louisville.edu.

REFERENCES

1. B. R. Brown and A. W. Lohmann, "Complex spatial filter," Appl. Opt. **5**, 967–969 (1966).
2. R. W. Cohn and L. G. Hassebrook, "Representations of fully complex functions on real-time spatial light modulators," in *Optical Information Processing*, F. T. S. Yu and S. Jutamulia, eds. (Cambridge U. Press, Cambridge, UK, 1998), Chap. 15, pp. 396–432.
3. N. C. Gallagher and B. Liu, "Method for computing kinoforms that reduces image reconstruction error," Appl. Opt. **12**, 2328–2335 (1973).
4. R. W. Cohn and M. Liang, "Approximating fully complex spatial modulation with pseudorandom phase-only modulation," Appl. Opt. **33**, 4406–4415 (1994).
5. R. W. Cohn and W. Liu, "Pseudorandom encoding of fully complex modulation to bi-amplitude phase modulators," in *Diffractive Optics and Microoptics*, Vol. 5 of 1996 OSA Technical Digest Series (Optical Society of America, Washington, D.C., 1996), pp. 237–240.
6. R. W. Cohn, "Pseudorandom encoding of complex-valued functions onto amplitude-coupled phase modulators," J. Opt. Soc. Am. A **15**, 868–883 (1998).
7. A. Papoulis, *Probability, Random Variables and Stochastic Process*, 3rd ed. (McGraw-Hill, New York, 1991), pp. 53–55 and 211–212.
8. R. W. Cohn and M. Liang, "Pseudorandom phase-only encoding of real-time spatial light modulators," Appl. Opt. **35**, 2488–2498 (1996).
9. R. W. Cohn, "Analyzing the encoding range of amplitude–phase coupled spatial light modulators," in *Spatial Light Modulators*, R. L. Sutherland, ed., Proc. SPIE **3297**, 122–128 (1998).
10. C. B. Burckhardt, "A simplification of Lee's method of generating holograms by computer," Appl. Opt. **9**, 1949 (1970).
11. U. Krackhardt, J. N. Mait, and N. Streibl, "Upper bound on the diffraction efficiency of phase-only fanout elements," Appl. Opt. **31**, 27–37 (1992).
12. L. B. Lesem, P. M. Hirsch, and J. A. Jordon, Jr., "The kinoform: a new wavefront reconstruction device," IBM J. Res. Dev. **13**, 150–155 (1969).
13. F. Wyrowski, "Upper bound of the diffraction efficiency of diffractive phase elements," Opt. Lett. **16**, 1915–1917 (1991).
14. R. D. Juday, "Optimal realizable filters and the minimum Euclidean distance principle," Appl. Opt. **32**, 5100–5111 (1993).
15. R. W. Cohn, A. A. Vasiliev, W. Liu, and D. L. Hill, "Fully complex diffractive optics by means of patterned diffuser arrays: encoding concept and implications for fabrication," J. Opt. Soc. Am. A **14**, 1110–1123 (1997).
16. A. Papoulis, *Probability, Random Variables and Stochastic Process*, 3rd ed. (McGraw-Hill, New York, 1991), pp. 43–56.
17. R. W. Cohn and M. Liang, "Spot array generator designed by the method of pseudorandom encoding," presented at the Annual Meeting of the Optical Society of America, Rochester, New York, October 20–24, 1996.
18. M. Duelli, D. L. Hill, and R. W. Cohn, "Frequency swept measurements of coherent diffraction patterns," Eng. Lab. Notes **9**, 3–5 (1998).
19. V. Kettunen, P. Vahimaa, and J. Turunen, "Zeroth-order coding of complex amplitude in two dimensions," J. Opt. Soc. Am. A **14**, 808–815 (1997).

Ternary pseudorandom encoding of Fourier transform holograms: errata

Robert W. Cohn and Markus Duelli

The ElectroOptics Research Institute, University of Louisville, Louisville, Kentucky 40292

[S0740-3232(99)01405-2]
OCIS codes: 230.6120, 090.1760, 030.6600, 070.0070.

Owing to the printing process the gray-level values below 60 (out of 256 levels) appear as black in Figs. 6 and 8 of Ref. 1. This makes it difficult in Fig. 8 to delineate between nonlinear effects of encoding and of the SLM. Figure 8 is reproduced here on a glossy paper and with the gray scale scaled by a factor of 2 (and also clipped for gray levels above 255). The inherent nonlinearity in nonrandom encoding produces large undesired diffraction orders that appear in the lower right corner of Fig. 8(a). Applying this encoding to a low-resolution phase-only light valve produces additional diffraction orders, including a bright spot on the optical axis and a set of orders at mirror locations to the desired spot array, as shown in Fig. 8(b). Applying ternary pseudorandom encoding to the same modulator produces the pattern in Fig. 8(c). This figure does not contain the undesired orders that are associated with the nonrandom algorithm of Fig. 8(a) but instead has a broadly spread, low-level background of speckle. Figure 6(i) in Ref. 1 is a closeup of Fig. 8(c). The speckle level is higher in Fig. 6(i) for three levels of quantization than in Fig. 6(h) for four levels of quantization. Figure 6(g) for five levels of quantization has an even lower level of speckle background. The corresponding Figs. 6(d)–6(i) for the simulated encodings show the same trends in background speckle levels. Readers who wish to view a version of Fig. 6 that has higher dynamic range can download the electronic version of the paper from *JOSA A* online or contact the authors for reprints.

The performance measures for phase-only nonrandom encoding were incorrectly reported in Table 3. The signal-to-noise ratio (SNR) was too small and the signal-to-peak-noise ratio (SPR) was too large. The correct

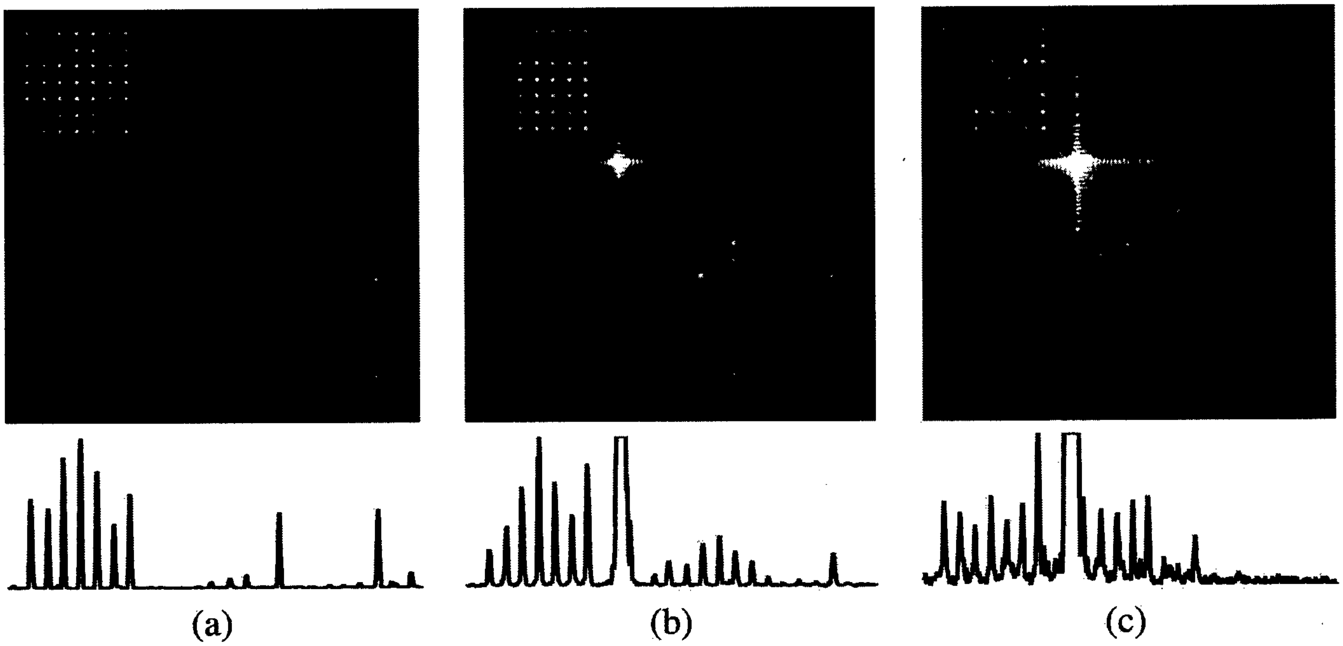


Fig. 8. Delineation of nonlinear effects on encoding: (a) simulated and (b) experimental diffraction pattern intensity for nonrandom ternary encoding, (c) experimental diffraction pattern for pseudorandom ternary encoding. These patterns show a larger view of the diffraction pattern than those in Figs. 6 and 7. Each intensity cross section is along the diagonal of the corresponding gray-scale image. In (a) and (b) the nonrandom ternary encoding produces mixing products, as evident in the lower left corner of each gray-scale image. Although speckle noise is evident in this same region for pseudorandom ternary encoding [(c)], it is much lower in intensity than the mixing products for (b). The saturated spot (centered on the optical axis) in (b) and (c) is primarily a result of the SLM cover glass not being antireflection coated. The most severe effect of the SLM's limited resolution is the appearance, to the lower left of the optical axis, of a duplicate 7 × 7 spot array in (b) and (c).

0740-3232/99/051089-02$15.00

numbers are SNR = 5400 and SPR = 17. This indicates that the phase-only pseudorandom encoding also produces a more faithful reconstruction than the nonrandom encoding since each encoding has identical SPR but the nonuniformity of the spot array for pseudorandom encoding is nearly half of that for nonrandom encoding.

Send all correspondence to Robert W. Cohn, The ElectroOptics Research Institute, Room 442, Lutz Building, University of Louisville, Lousiville, Kentucky 40292; tel, 502-852-7077; fax, 502-852-1577; e-mail, rwcohn01 @ulkyvm.louisville.edu.

REFERENCE

1. R. W. Cohn and M. Duelli, "Ternary pseudorandom encoding of Fourier transform holograms," J. Opt. Soc. Am. A. **16**, 71–84 (1999).

Pseudorandom encoding of complex-valued functions onto amplitude-coupled phase modulators

Robert W. Cohn

The Electro Optics Institute, University of Louisville, Louisville, Kentucky 40902-0001

Received April 30, 1997; revised manuscript received September 17, 1997; accepted October 31, 1997

Pseudorandom encoding is a method of statistically approximating desired complex values with those values that are achievable with a given spatial light modulator. Originally developed for phase-only modulators, pseudorandom encoding is extended to modulators for which amplitude is a function of phase. This is accomplished by transforming the phase statistics to compensate for the amplitude coupling. Example encoding formulas are derived, evaluated, and compared with a noncompensating pseudorandom-encoding algorithm. Compensating algorithms encode a smaller area of the complex plane and can produce more noise than is possible for arbitrary pseudorandom algorithms. However, the encoding formulas have greatly simplified numerical implementations. © 1998 Optical Society of America [S0740-3232(98)00704-2]

OCIS code: 050.1970.

1. INTRODUCTION

It is common to classify spatial light modulators (SLM's) as being either amplitude-only or phase-only, but, in practice, SLM's usually exhibit some degree of coupling between amplitude and phase (e.g., as illustrated in Fig. 1).[1] A notable example is liquid-crystal SLM's, which can be continuously changed from phase-mostly to amplitude-mostly operation by rotation of a wave plate or a polarizer.[2,3] Currently, no commonly available SLM's produce all complex values. Nonetheless, the design of diffractive optics and the implementation of other signal processing functions can often be simplified greatly if there are no constraints on the complex values. When the possible modulation values are constrained in some way, it has become common to employ numerically intensive global searches for functions that are implementable and that meet desired performance criteria. In many real-time applications using programmable modulators, these computational constraints may rule out the use of global searches. Although encoding does not usually match the performance of global searches, it can provide acceptable performance and numerically efficient and direct methods of representing fully complex functions with SLM's that are not fully complex.

The encoding problem considered in this paper is that of the design of Fourier transform holograms for implementation on available SLM's. One of the earliest discussions of this problem is by Brown and Lohmann.[4] Their methods use groups of pixels to represent a single complex value. Thus the space–bandwidth product of the SLM that uses such an algorithm will be reduced by the factor corresponding to the number of pixels in the group. Kirk and Jones introduced a point-oriented method of encoding complex values with a phase-only modulator.[5] The phase is specified to be the product of an amplitude-modulating function and a sinusoidal carrier. For discretely sampled phase-only SLM's, at least two pixels are required to represent one period of the sinusoid. Thus the space–bandwidth product of the signal is, at best, half that of the SLM. Cohn and Liang developed a method in which any desired complex values can be mapped to a single pixel, thereby using the entire space–bandwidth product of the SLM.[6] The method, referred to as pseudorandom encoding, uses ensemble averages of the values that are achievable with the SLM to represent the desired complex values. The actual modulation produced by the SLM corresponds to a single sample from the ensemble. The diffraction pattern of this random modulation consists of an approximate reconstruction of the desired diffraction pattern and a diffuse, approximately white-noise background. The method has also been interpreted as a carrier-based method.[7] Rather than using a single-frequency carrier, as does the Kirk–Jones method, a carrier of all frequencies is used. This diffracts unwanted light into all spatial frequencies. By a distribution of the unwanted light over the entire spatial bandwidth, the average noise level can often be much lower than the intensity of the desired reconstruction. This permits reconstructions to be formed anywhere over the bandwidth set by the grating frequency of the modulator.

Pseudorandom encoding also has many similarities with the parity sequence method of Chu and Goodman.[8] This method realizes a desired complex value by vectorial addition of two values of transmittance that are separated by $N/2$ pixels in an N-pixel phase-only modulator. This method perfectly reconstructs the desired diffraction pattern, but only at $N/2$ resolvable locations in the diffraction pattern. Between each sample of the desired reconstruction is a sample of the error signal (corresponding to the vector subtraction of the two values of transmittance), referred to as the parity sequence. Chu and Fienup described a version of this encoding method (named the synthetic coefficient method) in which the

0740-3232/98/040868-16$15.00

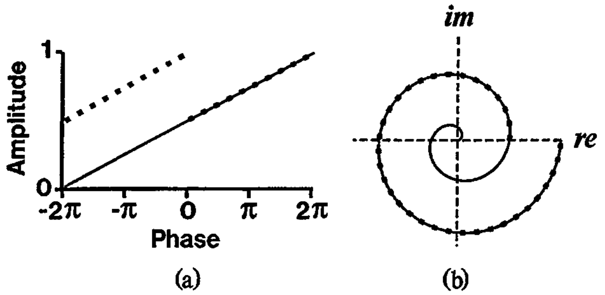


Fig. 1. Amplitude-coupled phase modulation characteristics illustrated in (a) rectangular and (b) polar plots. The characteristics have continuous phase ranges of 4π (solid curves) and 2π (dotted curves). Although amplitude coupling is drawn as a linear function, it is not necessary that it be linear for example 3 in Section 4.

transmittances of a pixel and its nearest neighbor are programmed as a group.[9] In this case the desired reconstruction is centered around the optical axis, and the error signal is centered at the Nyquist frequency $\pm N/2$. The desired reconstruction is most accurate on the optical axis, and its accuracy tends to decrease out to $\pm N/4$, where the error signal tends to dominate. Thus, as with any other group-oriented method, these methods, by their limiting the modulation bandwidth, are unable to use the entire space–bandwidth product of the SLM. However, in that pseudorandom encoding represents a desired complex value with only one pixel transmittance, it is possible to form a desirable reconstruction over the entire spatial bandwidth N for an N-pixel SLM. The error signal also reconstructs over this same spatial bandwidth. However, because pseudorandom encoding diffuses the error signal to (on average) a uniform level over the spatial bandwidth, it is often possible (depending on the desired complex function) for the desired reconstruction to be much brighter than the error. In that Chu and Fienup were studying the encoding of complex-valued functions, there are many other similarities between their methods and pseudorandom encoding. These include issues on trade-offs between diffraction efficiency and reconstruction accuracy, and the selection of phase degrees of freedoms. There are even some mathematical similarities between their encoding formulas and those for pseudorandom encoding. These points will be drawn out at appropriate places in the paper to distinguish the novel features of pseudorandom encoding.

Pseudorandom encoding was originally developed for phase-only SLM's that produce analog phase over a range of 2π[6] and, since then, for SLM's that produce analog phase and two levels of amplitude.[10] The phase-only encoding algorithms have been applied in laboratory demonstrations of beam shaping[11] and spot array generation.[7] This paper generalizes the pseudorandom-encoding concept for coupled modulators and describes new aspects of the encoding algorithms that appear when the modulation characteristics contain amplitude coupling. I will specifically consider that amplitude can be expressed as a function of phase. General derivations are presented together with concrete examples that use the two modulation characteristics illustrated in Fig. 1. Encoding for a variety of other characteristics can also be derived by the methods presented here. The use of these example characteristics is done only to provide consistency throughout the paper. A notable feature of the pseudorandom-encoding algorithms developed to date is that their implementation requires only a few numerical calculations per pixel. Thus, desired fully complex functions can be mapped to electrically addressable SLM's in real time by using low-end serial processors. A central goal of this investigation is to determine to what degree it is or would be possible to develop numerically efficient encoding algorithms for amplitude-coupled phase modulators.

Although the focus of this paper is on the design of encoding algorithms themselves, I have also included (in Section 8) a simulation of the diffraction patterns resulting from encoding one specific complex-valued function by various encoding algorithms. The results are used to illustrate the performance considerations of the algorithms that are discussed in Sections 5–7, as well as to provide comparisons with existing encoding algorithms (pseudorandom and others) that have already been developed for phase-only SLM's.

2. GENERAL DESCRIPTION OF PSEUDORANDOM-ENCODING AND PREVIOUS ENCODING METHODS

All pseudorandom-encoding algorithms specify the modulation of any given pixel in terms of a random variable. The statistical properties of the random variable are selected in such a way that the expected value, or average, of the random modulation is identical to the desired, but unobtainable, fully complex value. The desired complex-valued modulation is written as $\mathbf{a}_c = (a_c, \psi_c)$, and the resulting modulation by the SLM is $\mathbf{a} = (a, \psi)$, where the ordered pairs are the polar representations of the complex quantities. Complex quantities are indicated by boldface type. The pseudorandom-encoding design statement is, in general, to find a value of the ensemble average

$$\langle \mathbf{a} \rangle = \int \mathbf{a} p(\mathbf{a}) \mathrm{d}\mathbf{a} \tag{1}$$

of the random variable $\mathbf{a}$ such that $\langle \mathbf{a} \rangle = \mathbf{a}_c$. The statistical properties of $\mathbf{a}$ are determined by its probability density function (pdf) $p(\mathbf{a})$. The pdf is *selected* to ensure that the expected value of $\mathbf{a}$ and the desired complex value are identical. This selection of a pdf corresponds to solving the integral equation (1) for $p(\mathbf{a})$. [The solution is not unique, since the integral in Eq. (1) is a projection from the multidimensional space of $\mathbf{a}$ into a single value $\langle \mathbf{a} \rangle$. Various auxiliary conditions can be imposed on the solution and are considered in this paper.] After an appropriate density function is determined, the desired complex value $\mathbf{a}_c$ is encoded by drawing a single value of $\mathbf{a}$ from a random distribution having the density function $p(\mathbf{a})$. Since the value of $\mathbf{a}$ is found deterministically by computer, rather than from a random process occurring in nature, the procedure has been named pseudorandom encoding.

To this point the discussion has focused on encoding a single complex value. The procedure can be applied to encode spatially varying complex modulations $\mathbf{a}_c$. Specifically, in this paper I will assume that the SLM is a discretely sampled array of pixels. With the use of i as the spatial coordinate, the spatial samples of the desired complex modulation, the density function, and the random modulation are written as $\mathbf{a}_{ci}$, $p_i(\mathbf{a}_i)$, and $\mathbf{a}_i$. (This indexing scheme can be conveniently applied to one- or two-dimensional arrays, and it is not restricted to equally spaced samples.)

The far-field diffraction pattern of the encoded modulation $\mathbf{a}_i$ approximates the desired diffraction pattern. This can be seen by comparing the intensity of the desired far-field diffraction pattern with the ensemble average diffraction pattern that would result from the encoded modulation. The intensity pattern of the desired diffraction pattern is

$$I_c(f_x) = \left|\sum_i \mathbf{A}_{ci}\right|^2 = \left|\mathscr{F}\left\{\sum_i \mathbf{a}_{ci}\right\}\right|^2, \qquad (2)$$

where $\mathscr{F}\{\cdot\}$ is the Fourier transform operator; $\mathbf{A}_{ci}(f_x)$ is the Fourier transform of $\mathbf{a}_{ci}$, the desired complex transmittance of the ith pixel located at position i in the modulator plane; and f_x is the spatial coordinate across the Fourier plane. (Although this equation can also be written as a function of two spatial coordinates, one-dimensional coordinates are used throughout to simplify the presentation.) The expected value of the intensity pattern from the encoded modulation has been derived for the condition that the random variable $\mathbf{a}_i$ for the ith pixel is statistically independent of $\mathbf{a}_j$ for all j not equal to i. The ensemble average pattern is then expressed[6,10] as

$$\langle I(f_x)\rangle = I_c(f_x) + \sum_i (\langle|\mathbf{A}_i|^2\rangle - |\mathbf{A}_{ci}|^2), \qquad (3)$$

where $\mathbf{A}_i(f_x)$ is the Fourier transform of $\mathbf{a}_i$. The expected intensity consists of two terms. The first term is the desired diffraction pattern from Eq. (2). The second term represents the average level of background (i.e., speckle) noise that is produced as a result of the randomness of the modulation. For the case of pixels that are modeled as point sources, the average background noise is of constant intensity for all frequencies f_x (i.e., it is white). Many useful diffraction patterns can be synthesized for which I_c is accurately approximated and the noise level is adequately low.

The general encoding concept presented above makes no assumptions about the properties of the SLM or the specific statistical distributions selected. For coupled modulators in which amplitude $a(\psi)$ is a function of phase ψ, Eq. (1) becomes

$$\langle \mathbf{a}\rangle = \int a(\psi)p(\psi)\exp(j\psi)\mathrm{d}\psi \equiv a_0 \exp(j\psi_0), \qquad (4)$$

where $a_0 \equiv |\langle \mathbf{a}\rangle|$ is the effective amplitude and $\psi_0 \equiv \arg(\langle \mathbf{a}\rangle)$ is the effective phase resulting from the averaging operation. For the specific amplitude coupling $a(\psi) = 1$, Eq. (4) also describes the effective amplitude for phase-only modulators.

An important factor that controls the performance of not only pseudorandom encoding but several other encoding algorithms as well is the absolute magnitude scaling of the desired complex-valued function $\mathbf{a}_c$. I will use the symbol

$$\gamma = \max_i(\mathbf{a}_{ci}) \qquad (5)$$

for this scaling factor. For SLM's that are (usually assumed to be) passive devices, it may not be possible with some algorithms to encode complex values that exceed unity magnitude. As will be shown below, for some algorithms and modulation characteristics, the scaling parameter γ can be constrained to be much less than unity. Low values of γ reduce the amount of energy in the reconstruction and, even worse, can often increase the amount of noise energy. This observation is probably most easily seen for phase-only SLM's.[11] In this case all the energy incident on the modulator is transmitted to the Fourier plane. Thus, as γ decreases, the desired portion of the reconstruction, I_c, becomes increasingly dim and the noise component becomes increasingly bright. Consequently, the desired reconstruction becomes increasingly perturbed by noise and more difficult to see.

Chu and Fienup made quite similar observations for the parity sequence method.[9] Making $\gamma < 1$ ($A = 1/\gamma$ is used for the scaling parameter in their paper) reduces the energy in the desired reconstruction but does not affect accuracy. On the other hand, they also considered cases for $\gamma > 1$. For those desired magnitudes that exceed unity, the SLM transmittance is set to $\exp[j\arg(\mathbf{a}_{ci})]$, which they referred to as a kinoform but which today is more frequently referred to as a phase-only filter. They noted that it is possible to trade off the amount of energy in the diffraction pattern versus reconstruction accuracy for values of the scaling parameter, e.g., $1 \leqslant \gamma < \infty$. Similar sorts of trade-offs have been identified for some types of pseudorandom encoding in Refs. 7 and 10. In Sections 5–7 below, many new possibilities of blending together various pseudorandom (and also nonrandom) algorithms to improve the quality of the diffraction pattern are given. In Ref. 10 it has been shown that with blending it is even possible to obtain better reconstruction accuracy for $\gamma > 1$. Thus, from Section 5 on, I will frequently describe the encoding range of the various algorithms in terms of γ.

Closely related to γ is the diffraction efficiency of the *desired* (as opposed to the resulting) complex-valued function. Through the use of Parseval's relation, this can be calculated in the modulation plane as

$$\eta = \frac{1}{N}\sum_{i=1}^{N} |\mathbf{a}_{ci}|^2. \qquad (6)$$

This definition of diffraction efficiency has the physical interpretation of the average energy transmittance of the desired fully complex function. This result shows that for two different values of scaling factor, γ_1 and γ_2, the two resulting diffraction efficiencies vary according to $\eta_1/\eta_2 \propto (\gamma_1/\gamma_2)^2$. This relationship, together with Eq. (3), shows why (especially for pseudorandom encoding of phase-only SLM's) it is important to make the diffraction efficiency large. For the coupled SLM's considered in Fig. 1, there is a significant amount of phase modulation, and, as a result, making diffraction efficiency large usually leads to improved performance. (Simulations illustrating such improvements are presented in Section 8.)

The diffraction efficiency also depends on the desired function as well. This was recognized early in the development of phase-only computer-generated holograms and was also mentioned by Chu and Fienup[9] as well as Brown and Lohmann.[4] This has led to the continued development of code that optimizes the performance of the diffraction pattern under the condition that only the intensity of the diffraction pattern is of concern. The phases (the so-called degrees of freedom) are varied, with the goal of achieving near-unity diffraction efficiency and low reconstruction error. Because of the time-consuming nature of such optimizations, this type of design is not considered in this paper. Instead, a single complex-valued function is selected for the design examples in Section 8. Only the parameter γ is varied as permitted by the various encoding algorithms.

3. REVIEW OF PSEUDORANDOM ENCODING FOR PHASE-ONLY MODULATORS

This section reviews pseudorandom encoding for phase-only modulators and discusses the desirable features sought in developing a specific encoding formula. These results for phase-only encoding are used to motivate the derivations of encoding formulas in Section 4.

Various families of density functions $p(\psi;\ \langle\psi\rangle,\ \sigma)$, parameterized in terms of the mean value $\langle\psi\rangle$ and the standard deviation σ of the phase distribution, were evaluated in Eq. (4) of Ref. 6 and shown to produce all complex amplitudes having an amplitude between 0 and 1. In general, a two-dimensional search over $\langle\psi\rangle$ and σ is required to obtain a desired complex value. It was found (though not explicitly stated in Ref. 6) that the solution method could be simplified to a one-dimensional search by using density functions that are symmetric around their means $\langle\psi\rangle$. This led to the specific result that $\psi_0 = \langle\psi\rangle$ and that σ can be found by a one-dimensional search to obtain a desired value of effective amplitude in the range 0–1. Thus parameters that describe the density function (such as mean and variance) are individually associated with a value of effective phase and a value of effective amplitude. These conditions have led to simple expressions that can be evaluated with low numerical overhead for pseudorandom encoding of phase-only modulators. These points are brought out by way of the following example.

Example 1: Derivation of a pseudorandom phase-only encoding method. In Ref. 6 the effective complex amplitude was derived by using the uniform family of density functions [see Fig. 2(a)]

$$p(\psi;\ \langle\psi\rangle,\ \nu) = \frac{1}{\nu}\,\mathrm{rect}\left(\frac{\psi - \langle\psi\rangle}{\nu}\right), \tag{7}$$

which is specified in terms of the two parameters $\langle\psi\rangle$ and ν, the spread of the density function (where the spread is more conveniently used instead of the standard deviation $\sigma = \nu/\sqrt{12}$ in the case of uniform distributions). Substituting Eq. (7) into Eq. (4) for $a(\psi) = 1$ gives

$$\langle\mathbf{a}\rangle = \mathrm{sinc}(\nu/2\pi)\exp(j\langle\psi\rangle) \equiv a_0\exp(j\psi_0). \tag{8}$$

The amplitude is then identified as (solid curve in Fig. 3)

$$a_0 = |\mathrm{sinc}(\nu/2\pi)|, \tag{9}$$

and the phase is identified as

$$\psi_0 = \begin{cases} \langle\psi\rangle, & \mathrm{sinc}(\nu/2\pi) > 0 \\ \langle\psi\rangle + \pi, & \mathrm{sinc}(\nu/2\pi) < 0' \end{cases} \tag{10}$$

where the phase offset of π reflects sign changes of the sinc function. All values of effective amplitude between 0 and 1 can be realized by limiting the maximum spread to 2π. The effective complex amplitude then simplifies to

$$\langle\mathbf{a}\rangle = \mathrm{sinc}(\nu/2\pi)\exp(j\langle\psi\rangle),\quad 0 \leqslant \nu \leqslant 2\pi. \tag{11}$$

Equating Eq. (11) with the desired complex value $\mathbf{a}_c$, it is found that $\langle\psi\rangle = \psi_c$ and that

$$\nu = 2\pi\ \mathrm{sinc}^{-1}(a_c). \tag{12}$$

The desired value of $\mathbf{a}_c$ is encoded by selecting a random value of phase ψ from the distribution in Eq. (7) for the specified values of $\langle\psi\rangle$ and ν. In practice, this distribution is simulated by transforming the uniform random variable $s \in [0, 1]$, which has the pdf $p_s(s) = \mathrm{rect}(s - 1/2)$, into a random variable ψ that has the required

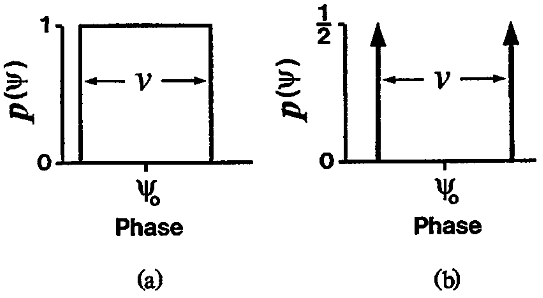


Fig. 2. Probability density functions (pdf's) for (a) uniform and (b) binomial random distributions of phase. In Section 4 effective pdf's of these same forms are sought.

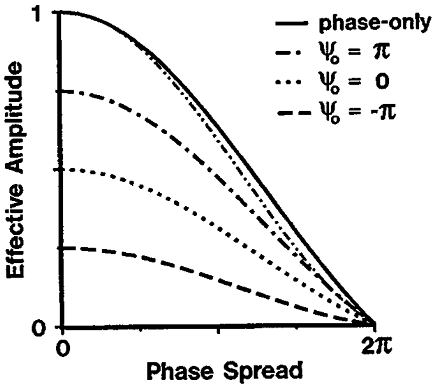


Fig. 3. Effective amplitudes for examples 1 and 2. The amplitude coupling used is identical to the solid curves in Fig. 1. The curve without a legend (dashed–dotted–dotted curve) is the effective amplitude for $\psi_0 = -\pi$ with the maximum effective amplitude normalized to unity. The two other curves even more closely match the sinc (phase-only) curve if they are normalized similarly.

pdf of Eq. (7). The random variable s is simulated by using the standard uniform random-number generator $s_i = \text{ran}(s_{i-1})$. The appropriate random value of phase is then drawn by using the formula

$$\psi = \langle \psi \rangle + \nu[\text{ran}(s) - 1/2]. \tag{13}$$

Equation (13) shows that pseudorandom encoding for phase-only modulation can be accomplished with a small number of mathematical operations. The speed of computation can be extremely fast if ν and ran(s) in Eqs. (12) and (13) are calculated with the use of lookup tables.

The encoding formulas (12) and (13) have a form that is quite similar to the parity sequence method of encoding. For both techniques there is a desired phase (i.e., $\langle \psi \rangle$ for pseudorandom encoding) and an effective-amplitude reduction (corresponding to, in the case of pseudorandom encoding, a random phase offset selected over a spread ν and, in the case of the parity sequence method, a fixed phase deviation between the two transmittances). In pseudorandom encoding, a spread of 2π produces an effective amplitude of zero [see Eq. (11)], whereas in Chu and Goodman's method a nonrandom spread of π effectively encodes a zero. An even closer mathematical correspondence is described for the algorithm derived in example 3 of Section 4.

4. DERIVATION OF PSEUDORANDOM-ENCODING ALGORITHMS FOR COUPLED MODULATORS

In this section the pseudorandom-encoding approach that was described in Section 3 is modified so as to compensate for the effects of amplitude coupling. Following this approach leads, once again, to encoding formulas that have low numerical overhead. The subsequent sections consider the range of realizable values (as measured by the value of γ) that can be encoded on coupled modulators by these algorithms and modifications of the algorithms that extend the encoding range.

The use of symmetric pdf's with amplitude-coupled modulation characteristics usually does not produce decoupling between the effective amplitude and the effective phase. The reason is that in Eq. (4) the coupled amplitude $a(\psi)$ is mathematically identical to a weighting function that biases the average of the random phasor $\exp(j\psi)$ away from $\exp(j\langle\psi\rangle)$. The bias is not usually constant, so that a two-dimensional search is required to find a solution to $(a_0, \psi_0) = (a_c, \psi_c)$. Alternatively, the form of the pdf can be chosen to compensate for amplitude coupling, so that it becomes possible to specify $\psi_0 = \psi_c$ directly. This is seen by defining the term $p_{\text{eff}}(\psi) \equiv a(\psi)p(\psi)$ in Eq. (4), which will be referred to as the effective pdf. Written this way, Eq. (4) has a form identical to that for the effective complex amplitude for phase-only encoding [Eq. (4) with $a(\psi) = 1$]. If the effective density function is symmetric about the amplitude-weighted expected value of phase

$$\langle \psi \rangle_{\text{eff}} = \int \psi p_{\text{eff}}(\psi) \mathrm{d}\psi, \tag{14}$$

then the effective phase ψ_0 will equal $\langle \psi \rangle_{\text{eff}}$ even though ψ_0 is not equal to the average phase $\langle \psi \rangle$. Thus the selection of the density function $p(\psi)$ that makes the effective density function $p_{\text{eff}}(\psi)$ symmetric permits the desired phase $\psi_c = \psi_0$ to be directly specified as the center of symmetry of the effective density function.

Although a desired value of ψ_0 can be directly specified, this approach does not entirely decouple a_0 from ψ_0. It does, however, produce the desired numerical simplification in that the effective amplitude a_0 becomes a one-dimensional function $a_0(\sigma; \psi_0)$ of σ (or ν) and the fixed parameter ψ_0. The desired value of amplitude corresponds to the value of σ that satisfies $a_0(\sigma; \psi_c) = a_c$. From the standpoint of numerical efficiency, sequential encoding of phase and amplitude is preferable to a simultaneous two-dimensional search for the encoding parameters.

The derivation of encoding formulas for coupled modulators also uses the following two results from probability[12]:

1. The effective pdf and the pdf $p(\psi)$ are not fully specified until a scale factor is determined that ensures that the integrated area of $p(\psi)$ equals unity. This requirement is simply part of the definition of a pdf. Specifically, the probability of the certain event is unity. This requirement can be expressed in terms of the cumulative pdf

$$P(\psi) = \int_{-\infty}^{\psi} p(\phi)\mathrm{d}\phi. \tag{15}$$

The random variable ψ has total probability $P(\psi) = 1$ for $\psi = \infty$.

2. Although random-number generators for arbitrary random distributions are not usually available, it is possible, by using a suitably chosen function, to transform the statistics of the uniform random variable s into the desired statistics. The function is known to be the inverse of the cumulative distribution function[12]:

$$\psi = P^{-1}(s). \tag{16}$$

The random variable ψ is then simulated by performing the function in Eq. (16) on the numbers produced by the random-number generator ran(s).

The procedure of deriving encoding formulas by this compensation approach is illustrated by the following two examples. In the first example (example 2), the amplitude-coupling function is given as an explicit function. In the second example (example 3), a solution is found in closed form without the amplitude-coupling function being given explicitly. This second form would be especially useful for SLM's for which the amplitude coupling can be changed *in situ* (for instance, liquid-crystal SLM's that are combined with rotatable polarizers or wave plates).

Example 2: Derivation for amplitude-coupling an explicit function. The amplitude-coupling function [solid line of Fig. 1(a)] is the linear function of phase

$$a(\psi) = m\psi + b, \qquad \psi \in [-2\pi, 2\pi], \tag{17}$$

where m is the slope and $b = a(0)$. A family of effective pdf's that is similar in form to Eq. (7) [Fig. 2(a)] is

$$p_{\text{eff}}(\psi) \propto \text{rect}[(\psi - \psi_0)/\nu]. \tag{18}$$

A single pdf is specified by two parameters: spread ν and bias ψ_0. After the correct normalizations are determined so that each pdf of the family $p(\psi;\ \psi_0,\ \nu)$ has unit area, the pdf that compensates $a(\psi)$ is identified as

$$p(\psi) = \frac{1}{\psi + b/m} \times \left[\ln\left(\frac{\psi_0 + \nu/2 + b/m}{\psi_0 - \nu/2 + b/m}\right)\right]^{-1} \operatorname{rect}\left(\frac{\psi - \psi_0}{\nu}\right). \tag{19}$$

With the use of Eqs. (15) and (16), the transformation from the uniform random variable s to the random variable ψ is found to be

$$\psi = \frac{(\psi_0 + \nu/2 + b/m)^s}{(\psi_0 - \nu/2 + b/m)^{s-1}} - b/m. \tag{20}$$

Substitution of Eqs. (17) and (19) into Eq. (4) gives a closed-form expression for the effective complex amplitude of

$$\langle \mathbf{a} \rangle = m\nu\left[\ln\left(\frac{\psi_0 + \nu/2 + b/m}{\psi_0 - \nu/2 + b/m}\right)\right]^{-1} \operatorname{sinc}\left(\frac{\nu}{2\pi}\right)\exp(j\psi_0). \tag{21}$$

Note the similarity between this equation and Eq. (11) for phase-only pseudorandom encoding. Furthermore, in the limit, as the slope m approaches zero, the effective complex amplitude for the coupled modulation approaches the effective complex amplitude for phase-only modulation.

The prescription for pseudorandom encoding, $\langle \mathbf{a} \rangle = \mathbf{a}_c$, and the relationship among $\langle \mathbf{a} \rangle$, ν, and ψ_0 in Eq. (21) can then be used to specify the following encoding algorithm:

```
select initial values for $s_0$, $b$, and $m$
For $i = 1$ to $N$ pixels
$\psi_{0i} \leftarrow \psi_{ci}$; $a_{0i} \leftarrow a_{ci}$; $s_i \leftarrow \text{ran}(s_{i-1})$
solve Eq. (21) for $\nu_i$ with $a_{0i}$ and $\psi_{0i}$ specified
calculate $\psi_i$ from Eq. (20) with $s_i$, $\psi_{0i}$, and $\nu_i$ specified.
```

To achieve the greatest computational speeds, the values of s_i, ν_i, and ψ_i would be precomputed and stored as lookup tables. Thus, by including amplitude compensation, it is possible to pseudorandom-encode coupled modulators in a manner similar to the encoding of phase-only modulators. The algorithms are similar in structure, and the encoding formulas are similar in form to those for pseudorandom phase-only encoding. The procedure illustrated in example 2 can be followed to derive encoding formulas for various other coupling functions and effective density functions.

In example 2 the availability of a 4π phase modulation range was assumed. This ensures that an effective amplitude between $a(\psi_0)$ (for a random phase spread $\nu = 0$) and zero ($\nu = 2\pi$) can be encoded for any value of ψ_0 over a 2π range. For the phase-only SLM the modulation characteristic is periodic, which eliminates the need for a SLM that produces phase in excess of 2π. Thus the algorithm proposed in example 2 suffers from the practical disadvantage that most SLM's available today barely produce a 2π phase range. A second disadvantage with the approach described in example 2 is that the algorithm needs to be custom designed for each individual modulation characteristic. Both disadvantages of the approach in example 2 are overcome in example 3 by designing a pseudorandom-encoding algorithm with the use of a different class of statistical distributions and by treating the amplitude modulation as a periodic function of phase.

Example 3: Derivation for amplitude coupling that is not an explicit function. The amplitude-coupling function is assumed be periodic. Figure 1(a) (dotted lines) illustrates a specific coupling function. The coupling is linear with phase, but no specific form is required or considered in this derivation. The discrete effective density function [see Fig. 2(b)]

$$p_{\text{eff}}(\psi) \propto \delta(\psi - \psi_0 + \nu/2) + \delta(\psi - \psi_0 - \nu/2), \tag{22}$$

where $\delta(\ \cdot\ )$, the Dirac delta function, is especially simple to use, since the amplitude-weighting compensation depends on only the amplitude values at the two points $\psi = \psi_0 \pm \nu/2$. Under these assumptions the analysis performed in example 2 can be repeated to give the pdf

$$p(\psi) = \frac{a(\psi_0 + \nu/2)\delta(\psi - \psi_0 + \nu/2) + a(\psi_0 - \nu/2)\delta(\psi - \psi_0 - \nu/2)}{a(\psi_0 - \nu/2) + a(\psi_0 + \nu/2)} \tag{23}$$

ρ and the effective complex amplitude

$$\mathbf{a}_c = \langle \mathbf{a} \rangle = \frac{2a(\psi_0 + \nu/2)a(\psi_0 - \nu/2)}{a(\psi_0 + \nu/2) + a(\psi_0 - \nu/2)} \cos(\nu/2)\exp(j\psi_0). \tag{24}$$

The ratio in Eq. (24) contains all the terms that compensate for the amplitude coupling. This term can be seen to be a ratio of the square of the geometric mean divided by the algebraic mean of the two amplitude samples $a(\psi_0 \pm \nu/2)$. This ratio reduces to a constant for phase-only SLM's. The uniform random variable s (from the available random-number generator) can be transformed into the random variable of phase by the simple threshold test

$$\psi = \begin{cases} \psi_0 - \nu/2 & \text{if } s \leqslant \dfrac{a(\psi_0 + \nu/2)}{a(\psi_0 - \nu/2) + a(\psi_0 + \nu/2)}. \\ \psi_0 + \nu/2 & \text{otherwise} \end{cases} \tag{25}$$

These results are especially useful in that this closed-form result applies to any function $a(\psi)$ for which ψ has a range of at least 2π. This formula can be applied even when samples are randomly selected from either side of an amplitude discontinuity [such as the discontinuity at $\psi = 0$ in Fig. 1(a)]. The encoding algorithm is identical to the algorithm following example 2 if Eqs. (24) and (25) are used in place of Eqs. (21) and (20), respectively.

The mathematics for this encoding algorithm are surprisingly similar to those for the parity sequence method.[8] This is most clearly seen for the case in which the coupling function $a(\psi) = 1$ for all values of ψ. The effective amplitude in Eq. (24) then becomes $a_c = \cos(\nu/2)$, and the effective phase is $\psi_c = \psi_0$. These effective values are identical to those that are encoded by the parity sequence method. Furthermore, the actual modulation produced by the pseudorandomly encoded phase-only modulator is either $\mathbf{a} = \exp[j(\psi + \nu/2)]$ or $\mathbf{a} = \exp[j(\psi - \nu/2)]$ with probability 1/2. The only difference with the parity sequence method is that each of these transmittances is produced by a pair of spatially separated pixels. In general, when there is coupling, the probability of selecting one or the other value of $\mathbf{a}$ is adjusted [(according to Eq. (25)] so that the weaker modulation is selected more frequently than the stronger modulation. This compensation effectively attenuates the stronger modulation, so that both values of modulation are effectively of equal strength.

5. NUMERICAL EVALUATION OF THE ENCODING FORMULAS AND IMPLEMENTATION ISSUES

The characteristics and the implementation of the encoding formulas can be appreciated by considering some numerical simulations of effective amplitude. The amplitude-coupling functions shown in Fig. 1 will be used. This corresponds to using a slope of $m = 1/4\pi$ and $b = 1/2$ in the equations in example 2. The effective amplitudes found from Eq. (21) are plotted in Fig. 3. In example 3 the amplitude-coupling function used is similar to Eq. (17). A slope of $m = 1/4\pi$, and $b = 1/2$ for $\psi > 0$ and $b = 1$ for $\psi < 0$, are used to describe the two line segments [the dotted lines in Fig. 1(a)]. Using this coupling function in Eq. (24) produces the effective amplitudes shown in Fig. 4.

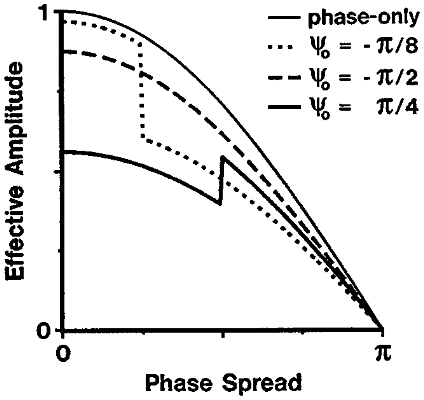


Fig. 4. Effective amplitude curves for example 3. The amplitude coupling is identical to the dotted curve in Fig. 1(b). The effective amplitudes for phase-only encoding with use of the pdf from Fig. 2(b) are included for comparison.

Numerical results for example 2. Figure 3 shows that the range of possible effective amplitudes a_0 depends on the value of the effective phase ψ_0. For a spread of $\nu = 0$, the effective amplitude of each curve is identical to $a(\psi_0)$, and as the spread increases, the effective amplitude decreases monotonically from this point. The shape of each curve is very close to that of a sinc function, as is illustrated by the normalized version of the effective amplitude for the $\psi_0 = -\pi$ curve. The two other effective amplitudes are even closer in shape to that of a sinc. Numerically efficient solutions of ν that are only slightly more involved than those for phase-only encoding can be developed by using this result.

The curves plotted in Fig. 3 correspond to those for which ψ_0 is contained between $-\pi$ and π. For this range of effective phase, all the curves range between $a(\psi_0)$ and 0. For $|\psi_0| > \pi$ the 4π range of the modulator limits the maximum spread to $2(2\pi - |\psi_0|)$. Thus, for these particular values of ψ_0, the spread cannot be made large enough to continuously reduce a curve of effective amplitude to 0. In this subsection I will consider the implications of using only the effective-amplitude curves for which $-\pi \leq \psi_0 \leq \pi$, and I will consider the more general case, for which $-2\pi \leq \psi_0 \leq 2\pi$, in Section 6.

If one chooses to use the Fig. 3 curves to pseudorandom-encode the coupled modulator, the desired complex function $\mathbf{a}_c$ must be scaled by an appropriate choice of γ [see Eq. (5)]. Consider that if $\gamma > 0.25$, then amplitudes for some values of phase cannot be encoded. However, if $\gamma = 0.25$, then any complex value out to a circular radius of γ can be pseudorandom encoded. Since γ can be as large as 1 for pseudorandom encoding on a phase-only modulator, it can be seen that the diffraction efficiency η of the coupled modulator will be 1/16 of the efficiency of the phase-only modulator [see Eq. (6)]. Of even more concern than efficiency is that by scaling the maximum magnitude to $\gamma = 0.25$, large random phase spreads are needed to encode many of the desired complex values. For instance, for the $\psi_0 = \pi$ curve in Fig. 3, all spreads are between approximately 1.5π and 2π. Thus substantial amounts of random noise can be generated by applying this particular algorithm for a SLM that has the characteristic of Fig. 1. This observation has motivated the development of various modified encoding algorithms (presented below) for which γ can be larger. Of course, if the amplitude coupling is not as strong as that considered in this example, then the value of γ can be made correspondingly larger. Certainly, such an algorithm would be well suited to SLM's that are phase mostly (amplitude variation between 0.9 and 1).[2,3] The example 2 encoding algorithm appears to adapt phase-only encoding better to phase-mostly SLM's than to SLM's for which phase is strongly coupled to amplitude.

There are other options for selection of the amplitude degree of freedom. Consider (for the Fig. 3 results) that γ is between 0.25 and 0.75. For the complex values that cannot be pseudorandom encoded, one or more additional encoding methods would be combined with the pseudorandom-encoding formula. One appropriate encoding technique is the deterministic mapping method that maps complex values to the closest point on the modulation curve.[13] Earlier studies on encoding 'complex-valued composite functions onto phase-only[14] and biamplitude phase[10] modulators have already demonstrated that the blending of deterministic and pseudorandom encoding can produce performance that is better than either. The numerical implementation of blended encoding algorithms is not substantially more involved. However, it is more numerically intensive, since, at present, the only way to specify the scaling of the desired complex values that produces the best encoding performance is to perform the encoding repetitively for different scaling factors and then evaluate the performance of the resulting modulation for each encoding.

Summarizing the Fig. 3 results: It has been shown that the encoding of weakly coupled (phase-mostly) SLM's can be performed with a small amount of additional numerical overhead and a slight reduction in the performance as compared with that of phase-only pseudorandom encoding. For strongly coupled SLM's the reduction in performance can be significant and may call for the development of more numerically involved algorithms that blend individual pseudorandom-encoding algorithms with other encoding algorithms.

Numerical results for example 3. The evaluation of the effective amplitude shows that there are three types of effective-amplitude curves. These are for (I) $\pi/2 \leqslant \psi_0 \leqslant 3\pi/2$, (II) $0 \leqslant \psi_0 \leqslant \pi/2$, and (III) $3\pi/2 \leqslant \psi_0 \leqslant 2\pi$. Similar to the results of example 2, the type I curves descend monotonically from $a(\psi_0)$ to 0 as illustrated in Fig. 4 by the curve $\psi_0 = -\pi/2$. (Because of the periodic assumption, this is also referred to as the $\psi_0 = 3\pi/2$ curve.) With the use of a binomial distribution, zero amplitude is realized for a spread of π as opposed to 2π in example 2. For type II curves (e.g., the $\psi_0 = \pi/4$ curve in Fig. 4), a discontinuity in the effective amplitude is found for spread $\nu = 2\psi_0$. This is due to the discontinuous jump between 0.5 and 1 in the value of the term $a(\psi_0 - \nu/2)$ in Eq. (22). The jump produces a range of values for which there are two solutions for a desired effective amplitude. For some values of ψ_0, the effective amplitude can be even larger than the effective amplitude for zero spread. For the type III curves (e.g., the $\psi_0 = -\pi/8$ curve in Fig. 4), there is a discontinuity at $\nu = -2\psi_0$ [for ψ_0 expressed as a negative angle or $\nu = 2(2\pi - \psi_0)$ for ψ_0 expressed as a phase between π and 2π]. The discontinuity in effective amplitude for any given curve shows that some values between $a(\psi_0)$ and 0 cannot be encoded by the formulas in example 3.

The region of the unit disk for which the desired complex values can be encoded is shown in Fig. 5. Complex values in the clear region can be pseudorandom encoded, and values in the striped region cannot. The dotted curves represent the values to each side of the discontinuity for the type II curves. Note that, for ψ_0 between 0 and slightly less than $\pi/4$, effective amplitudes greater than $a(\psi_0)$ can be encoded. The portion of the striped region forming a peninsula that spirals into the origin corresponds to the unrealizable complex values for the type III curves. Its boundaries correspond to the values on each side of the discontinuity of the type III curves.

For the 4π modulator (with use of the example 3 algorithm), it is possible to scale the complex values so that all values can be pseudorandom encoded. For the results in this subsection for the 2π modulator, there is no value of γ that will permit the pseudorandom encoding of all desired complex values. This does not mean that the example 3 formulation is less desirable than the formulation in example 2. One advantage of the example 3 method is that it requires a 2π, as opposed to a 4π, modulator. A second advantage is that it encodes a much greater area of the unit disk than the example 2 method. Blending with the deterministic encoding algorithms, as described above, can be used to encode values in the striped peninsular region of Fig. 5. As discussed in Section 2, it is even possible that using deterministic algorithms to encode values outside the unit disk (i.e., $\gamma > 1$) can sometimes produce better performance. Thus deterministic encoding could be used to realize all values on the striped portion of the unit disk, in addition to the complex values that have amplitudes greater than unity.

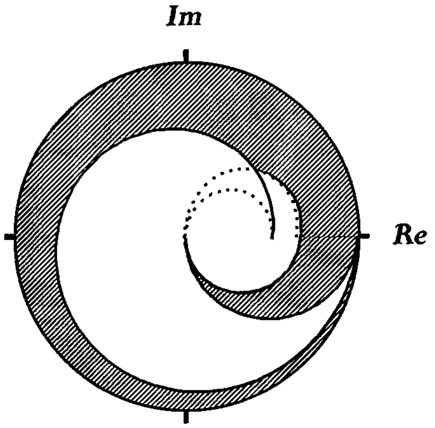


Fig. 5. Map of the effective complex amplitudes that can be pseudorandom encoded (clear region) and those values that cannot be pseudorandom encoded (striped region) on the unit disk by the encoding method from example 3. The amplitude coupling used is the same as the dotted curve in Fig. 1(b), and it is replotted in this illustration. Note that for effective phases between 0 and ~$\pi/4$ the effective amplitude can be larger than the amplitude of the coupling function. For effective phases between $-\pi/2$ and $\pi/2$, the effective-amplitude curves (Fig. 4) have jump discontinuities. For effective phases between 0 and $\pi/2$, the values on each side of the discontinuity are plotted as dotted curves. For effective phases between $-\pi/2$ and 0, the discontinuities form the boundary of the portion of the unrealizable region that has radii less than the amplitude-coupling function.

From the results in this section, it can be seen that some complex values that are encoded by one pseudorandom algorithm may not be encoded by another. This is due to the specification of the form of the pdf's in the derivation of the encoding formulas. Thus, rather than using a deterministic algorithm for these values, it is possible to increase the area of the unit disk that can be encoded by combining two pseudorandom-encoding algorithms. However, as mentioned in Section 2, there are an uncountable number of pdf's that satisfy Eq. (1). This raises the question: What is the total extent of the complex plane that can be encoded by all possible pseudorandom-encoding algorithms? This is answered in Section 6, which also contains a discussion of various approaches for modifying the pseudorandom-encoding algorithms to increase their range.

6. EXTENDING THE RANGE OF PSEUDORANDOM ENCODING

Four methods of extending the complex range are described. For the sake of clarity, the discussion pertains specifically to the examples presented herein, even though the results can be similarly applied to other coupling functions and pdfs. The four methods are (1) for the 4π modulator, using all effective values of phase ψ_0 between -2π and 2π for encoding; (2) for the 4π modulator, blending the example 2 and example 3 pseudorandom algorithms; (3) for either modulator, randomly combining two pseudorandomly encoded values to encode a previously unrealizable value; and (4) for the 2π modulator, using a pseudorandom encoding that does not compensate for amplitude coupling. This last method is also used to evaluate which values can and cannot be implemented by any possible means of pseudorandom encoding.

Method 1. The example 2 pseudorandom encoding algorithm has been applied above to realizing complex amplitudes for effective phases $-\pi \leq \psi_0 \leq \pi$. The rationale for this is that spread ν can then be varied between 0 and 2π, which allows all values of effective amplitudes between $a(\psi_0)$ and 0 to be encoded for any given effective-amplitude curve (e.g., those shown in Fig. 3). There is a substantially greater area of the complex plane that can be pseudorandomly encoded if the effective complex amplitudes for $\pi \leq \psi_0 \leq 2\pi$ are also admitted. (Additional solutions for effective phase $-2\pi \leq \psi_0 \leq -\pi$ are also possible but are not considered here, since these solutions do not increase the area that can be pseudorandomly encoded.) Over this extended range of effective phase, the spread ν can be varied from 0 to a maximum of $2(2\pi - \psi_0)$. The minimum values of effective amplitude calculated by using Eq. (21) are plotted (dashed curve) in Fig. 6. Figure 6 shows that a substantially greater area of the complex plane can be encoded than had originally been considered for the example 2 method. Previously, the complex values had been scaled to a maximum amplitude of $\gamma = 0.25$ to use the pseudorandom-encoding algorithm by itself. The point where the dashed curve crosses the effective-amplitude curve $a(\psi)$ is $\psi_0 = 4\pi/3$. Thus the desired amplitudes can be normalized to $\gamma = a(4\pi/3) = 0.33$ instead of $\gamma = 0.25$. Also, a large area (the clear peninsular region) is available for pseudorandom encoding if $\gamma > 0.33$ is used.

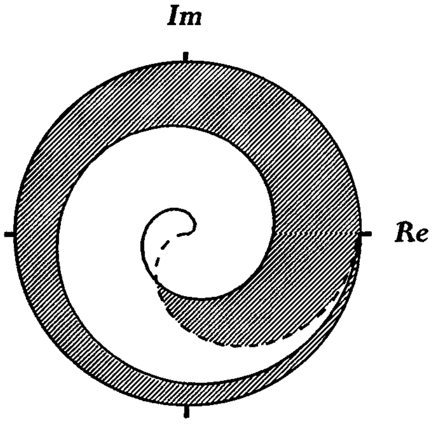


Fig. 6. Map of the effective complex amplitudes that can be pseudorandom encoded (clear region) and those values that cannot be pseudorandom encoded (striped region) on the unit disk by the encoding method from example 2. Additionally, the minimum value of effective amplitude has been calculated for phases between π and 2π by using Eq. (21) (dashed curve). This has increased greatly the number of complex values that can be encoded. The amplitude coupling used is the same as the solid curve in Fig. 1(b), and it is replotted in this illustration.

Method 2. Although neither the example 2 nor the example 3 encoding algorithm is capable of encoding all complex values, each can encode values that the other cannot. Even though the unrealizable area for the example 2 algorithm is larger than that for the example 3 algorithm, the two algorithms taken together encode an even larger area than the two taken separately. By overlaying Figs. 5 and 6, it can be seen that all complex values having amplitudes of $\gamma = 0.43$ or less can be encoded by combining the two algorithms.

Method 3. Two effective complex amplitudes may be pseudorandomly combined to realize values in the peninsular regions in Figs. 5 and 6. I will show this for the specific condition that both effective amplitudes have the same effective phase ψ_0. This is in keeping with the goal throughout this paper of producing encoding formulas for which a range of effective amplitudes can be specified as a function of spread for any given value of effective phase. In general, the new effective complex amplitude can be written as

$$\langle \mathbf{a} \rangle = d\langle \mathbf{a} \rangle_{\mathbf{u}} + (1-d)\langle \mathbf{a} \rangle_{\mathbf{l}}, \tag{26}$$

where the subscripts **u** and **l** are used to distinguish the upper/larger and lower/smaller effective complex amplitudes. These values could be selected to correspond to the complex amplitudes to each side of the peninsular region at a given effective phase. The value d is the probability of randomly encoding the upper value, and $1-d$ is the probability of encoding the lower value. Under this set of assumptions, the effective complex amplitude is written as

$$\langle \mathbf{a} \rangle = [da_u + (1-d)a_l]\exp(j\psi_0), \tag{27}$$

where a_u and a_l are the individual effective amplitudes. The effective amplitude a_0 for Eq. (27) can be seen to produce any amplitude between a_u and a_l for values of d between 0 and 1. This shows that the method could be used to encode the values in the peninsular region. This method generalizes the biamplitude phase encoding algorithm in Ref. 10 by using effective amplitudes instead of amplitudes from the modulation characteristic. Method 3 can also be viewed as encoding with the use of a different density function. This can be shown by explicitly expressing the effective amplitudes in Eq. (27) as integrals with the use of Eq. (1). The arguments of the integrals can be collected to form a single integral. This permits the identification of the single pdf

$$p(\psi;\, d) = dp_u(\psi;\, \psi_0,\, \nu_u) + (1-d)p_l(\psi;\, \psi_0,\, \nu_l), \tag{28}$$

where ν_u and ν_l are the two values of spread and p_u and p_l are used to distinguish the two density functions and the new density function p from each other. Although the pdf p is shown to depend on only the parameter d, it also depends on ν_u, ν_l, and ψ_0. These dependencies are not shown because in this encoding method the effective complex amplitudes $\langle \mathbf{a} \rangle_\mathrm{u}$ and $\langle \mathbf{a} \rangle_\mathrm{l}$ are determined individually, which completely determines p_u and p_l. Then only d needs to be specified to complete the encoding.

If this extension is used, it becomes possible to pseudorandom-encode the entire peninsular regions, in addition to the clear areas in Figs. 5 and 6. If this method is used with the example 2 algorithm, then all complex values of amplitude $\gamma = 0.5$ or less can be encoded by pseudorandom encoding alone. If the example 3 algorithm is used instead, then fully pseudorandom encoding can be used if the complex values are normalized to a maximum amplitude of approximately $\gamma = 0.56$ (on account of the presence of the clear peninsular region in Fig. 5 that covers the phase range of 0 to approximately $\pi/4$).

Method 4. The pseudorandom-encoding range can be extended further by using different pdf's in deriving the encoding formulas. The realizable range can be evaluated by using, as in example 3, a binomial distribution. However, in this evaluation the distribution is not constrained to compensate for bias drift of the effective phase, and the spread is not constrained to be symmetric about the value of effective phase. The evaluation follows from considering the effective complex values that can be encoded by a given pair of complex values $\mathbf{a}_1$ and $\mathbf{a}_2$ on the modulation characteristic. Using the binomial density function in Eq. (4) gives an expression for the effective complex amplitude of

$$\langle \mathbf{a} \rangle = d\mathbf{a}_1 + (1 - d)\mathbf{a}_2, \tag{29}$$

where d is the probability of selecting $\mathbf{a}_1$. The expression is identical to Eq. (26), except that the sample points are constrained to lie on the modulation characteristic. Equation (29) is recognized as the expression for a line as a function of the variable d. Thus any value lying on the line segment between $\mathbf{a}_1$ and $\mathbf{a}_2$ can be encoded. This is illustrated in Fig. 7 for four pairs of points. The three line segments drawn as thin lines all cross at a common point. Thus any one of these segments, or an infinite number of other line segments, could be used to encode this particular complex value. This observation is equally valid for any desired complex value found in the clear region in Fig. 7 that is bounded by the modulation characteristic curve (dashed curve) and the thick line segment. The interior of this boundary is a convex set of all the complex values that can be encoded by the union of all possible pseudorandom-encoding algorithms. Values on the boundary are also realizable, but they have only a single possible solution. Values outside the boundary cannot be realized because the ensemble average of a random phasor never produces a magnitude that is larger than the average magnitude of the phasors in the ensemble.

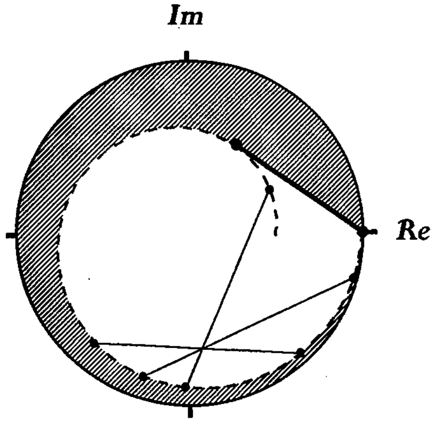


Fig. 7. Map of the effective complex amplitudes that can be pseudorandom encoded (clear region) and those values that cannot be pseudorandom encoded (striped region) on the unit disk by use of all possible binomial distributions. The amplitude coupling (dashed curve) is the same as the dotted curve in Fig. 1 (b). For a given pair of samples on the amplitude-coupling curve, any effective complex amplitude can be realized on the line segment connecting the two points. This is shown for four pairs of samples. Any of the three thin lines could be used to produce the same effective complex amplitude at their common intersection. The thick line segment and the amplitude-coupling curve bound a convex set of all complex values that can be realized by pseudorandom encoding for the particular modulator characteristic.

This simple evaluation is also useful for evaluating the range over which pseudorandom encoding can be applied to modulators that do not produce a full 2π range of phase modulation. For the specific curve and construction in Fig. 7, pseudorandom encoding can be used to encode fully complex functions of any phase and amplitude less than approximately $\gamma = 0.58$. This means that the same function encoded by method 4 would have [by Eqs. (5) and (6)] a diffraction efficiency 5.4 times greater than the example 2 algorithm for which $\gamma = 0.25$. The efficiency of method 4 would also be one third of that for pseudorandom phase-only encoding where $\gamma = 1$.

These results answer the question about the maximum range possible. They also raise a new question about which of the possible solutions to the encoding problem is preferable. The approaches considered to this point emphasize the finding of formulas that are simple to implement and that are numerically efficient. For the example 3 algorithm, in which amplitude compensation was used to simplify the implementation, no more than two binomial distributions (i.e., line segments) could be found to realize a desired complex value, and for some desired and realizable complex values no distribution could be found. This section shows that the limitations introduced by these particular assumptions can be substantially reduced by the use of Methods 1–3. The benefit of simple implementations is that they provide great flexibility and easy access to the complex modulating properties of a SLM in an environment that can require real-time programming of the SLM. If the highest levels of optical performance are required and computation time is not a significant concern, then there are many numerically intensive design algorithms that can produce near-optimal optical performance. These considerations have led to my emphasis on simply implemented algorithms. However, it would be possible instead to derive pseudorandom-encoding algorithms on the basis of minimizing the random errors produced by encoding itself. This is the subject of Section 7. The resulting algorithms appear to be more numerically involved, and, for this reason, it is not my intent to recommend them. Instead, the minimum-error (and also the maximum-error) pseudorandom encoding formulas are used to bound the errors produced by amplitude-compensated algorithms.

7. ERROR ANALYSIS OF PSEUDORANDOM ENCODING AND MINIMUM-ERROR ENCODINGS

Encoding error defined. A general error analysis of pseudorandom encoding, not only for phase-only but also for any SLM modulation characteristic is presented in Ref. 6. Expressions are presented that, in addition to describing the expected intensity, describe the standard deviation of the intensity pattern. For the discussions in this section, I evaluate only the expected intensity, which is given in Eq. (3). The noise, or encoding error, has been identified in Section 2 as the second term in Eq. (3). The ith element of this summation corresponds to the encoding error produced by the ith SLM pixel. The evaluation is most directly performed and the key results are most readily apparent under the assumption that the pixel aperture is an ideal impulse (Ref. 6 can be consulted for further analyses that include finite-width apertures). For this set of assumptions, the error for a single pseudorandom-encoded pixel is

$$\epsilon = \int a^2(\psi)p(\psi)\mathrm{d}\psi - |\mathbf{a}_c|^2. \qquad (30)$$

Since the desired complex amplitude is fixed, the integral is the only term that can be minimized. The integral is minimized by individually minimizing the integrand at each value of ψ. Since $a(\psi)$ is a given function, the minimum is produced by selecting $p(\psi)$ to be minimum where $a(\psi)$ is maximum. The minimization is subject to the two constraints that (1) the integrated area of the density function is unity and (2) the value $\mathbf{a}_c = \langle \mathbf{a} \rangle$ in Eq. (4) has a solution. The generality of Eq. (30), together with the constraints, makes it difficult to draw additional conclusions about minimum-error pseudorandom encodings without further specialization of the problem.

Consider the specific problem of determining which binomial distribution produces the least encoding error (as illustrated in Fig. 7). Under this set of constraints, the encoding error in Eq. (30) can be expressed as

$$\epsilon = d(1-d)[a_1{}^2 + a_2{}^2 - 2a_1a_2\cos(\psi_1 - \psi_2)]$$
$$\text{subject to } \mathbf{a}_c = d\mathbf{a}_1 + (1-d)\mathbf{a}_2, \qquad (31)$$

where $\mathbf{a}_i \equiv (a_i, \psi_i)$. The error can be minimized by separately minimizing the product of the binomial probabilities and minimizing the term in brackets. The product $d(1-d)$ is minimized by maximizing the distance between d and $1-d$. The term in brackets is the familiar formula for the law of cosines. It is minimized by minimizing the difference between the two phases. The constraint in Eq. (31) still makes it difficult to draw a useful conclusion. If this constraint is substituted into the encoding error in Eq. (31) to eliminate the value of probability d, then this equation reduces to

$$\epsilon = |\mathbf{a}_c - \mathbf{a}_1||\mathbf{a}_c - \mathbf{a}_2|. \qquad (32)$$

Written in this form, the pseudorandom-encoding error can be directly interpreted as the product of the lengths of the line segments $\mathbf{a}_1 - \mathbf{a}_c$ and $\mathbf{a}_c - \mathbf{a}_2$. For a fixed length segment $\mathbf{a}_1 - \mathbf{a}_2$, the product is minimized by making one of the two segments as large (or as small) as possible. However, the length of $\mathbf{a}_1 - \mathbf{a}_2$ generally varies, depending on the form of the amplitude-coupling function $a(\psi)$. Thus further analysis is usually required to find the solution that produces a minimum error.

Numerical analysis of encoding errors for binomial distributions. Minimum-error encoding formulas have been determined numerically for a number of desired complex values for the modulation characteristic in Fig. 7. The method of finding the minimum-error encoding formula for a single complex value consists of repeated evaluations of the error ϵ by Eq. (32). The steps are the following: (1) A value of $\mathbf{a}_c$ is specified; (2) a value of phase ψ_1 is specified; (3) the complex amplitude $\mathbf{a}_1 = a(\psi_1)\exp(j\psi_1)$ is calculated; (4) $\mathbf{a}_2$ is found by solving for the point on the line through $\mathbf{a}_1$ and $\mathbf{a}_c$ that intersects $\mathbf{a}_2(\psi_1) = a(\psi_2)\exp(j\psi_2)$; (5) the error is found by using Eq. (32). Steps 2–5 are repeated for all values of ψ_1. From these results the values of ψ_1 and ψ_2 are determined that produce the minimum error for the encoding of the value $\mathbf{a}_c$. Step 4 requires the solution of the nonlinear equation $\mathbf{a}_2(\psi_2) = \mathbf{a}_1 + x(\mathbf{a}_c - \mathbf{a}_1)$. Separately equating the real and imaginary parts of this expression gives two equations in the two unknowns x and ψ_2. These values were solved by using the "find" function in the software package MATHCAD (Mathsoft Inc., Cambridge, Mass. 02142).

The complex value $\mathbf{a}_c = (0.67, 1.6\pi)$ corresponds to the intersection of the three thin line segments in Fig. 7. The minimum-error encoding formula corresponds to a line segment that contains $\mathbf{a}_c$ and the sample point at (0.5, 0). In fact, for this particular coupling characteristic, one of the two sample points is usually (0.5, 0). In all other cases one of the two points is $(1, 2\pi)$. This corresponds to the encoding of those effective amplitudes that exceed $a(\psi_0)$ (i.e., for effective phase between 0 and approximately 0.32π). This procedure can be summarized in the following way: Pick the amplitude of one of the sample points to be as small as possible. This is true even if one sample point is $(1, 2\pi)$. In this case the other sample point is the minimum possible.

Comparison of minimum-error and amplitude-compensated encoding algorithms. The encoding error produced by the example 3 algorithm can be evaluated by solving Eq. (24) for the spread ν for which $\mathbf{a}_c = (a_0, \psi_0)$. Then the complex values on the modulation characteristic, $\mathbf{a}_1$ and $\mathbf{a}_2$, are calculated for the phases $\psi_0 \pm \nu/2$. Using these two sample values in Eq. (32) gives the encoding error ϵ_c for encoding the value $\mathbf{a}_c$ by the compensation method. These errors are compared with minimum encoding error $\epsilon_{\min}$ in Table 1 for selected values of $\mathbf{a}_c$. Errors up to three times larger than the minimum error are produced by the compensation method. However, individual errors can be much closer in value to the minimum error. For an effective amplitude of 0.68 in the table, the compensated encoding has a spread of $\nu = \pi/10$. Since one sample point is $(1, 2\pi)$, the compensated encoding coincides with the minimum-error encoding. The additional error produced by the compensation method can be appreciated by considering the errors produced for $a_0 = 0$. The phase of a desired complex value need not be specified for zero-valued amplitudes, but, depending on the specification of ψ_0, the encoding error varies from 0.375 for the sampling points (0.5, 0) and $(0.75, \pi)$ to 0.75 for the sampling points $(0.75, \pi)$ and $(1, 2\pi)$.

The search procedure used to identify the minimum encoding error has also been used to identify the maximum possible encoding error $\epsilon_{\max}$. The maximum-error solution usually corresponds to one of the sample points being $(1, 2\pi)$. For the case $a_0 = 0.68$ in Table 1, for which the minimum-error solution uses the $(1, 2\pi)$ sample, the maximum-error solution is found by selecting the smaller of the two sample points to have the smallest amplitude possible. This corresponds to the line segment intersecting $\mathbf{a}_c$ that is tangent to the modulation characteristic. In Table 1 the encoding error for the compensated algorithm, ϵ_c, approaches the upper limit $\epsilon_{\max}$ closely for $\psi_0 = 1.6\pi$ and 1.75π. The solution for the compensated encoding of these two points includes one sample point that is close to $(1, 2\pi)$, whereas a sample point at (0.5, 0) is needed for minimum error.

For comparison I also include in Table 1 the encoding errors for a phase-only SLM. As with phase-only modulation in general, for pseudorandom phase-only modulations there is a conservation of the energy diffracted from the modulation to the Fraunhofer plane.[6] For this reason the encoding error for phase-only modulation is $\epsilon_{\mathrm{po}} = 1 - {a_0}^2$. (This result is independent of the particular phase-only pseudorandom-encoding algorithm. Also note that the square root of this error is identical to the magnitude of the error vector of the parity sequence method.[8]) Since the phase-only characteristic has a greater radius than that of the coupled curve, it is not surprising that the errors for the phase-only characteristic are larger than those for $\epsilon_{\max}$.

Consideration of other coupling functions. It is not immediately obvious if the minimum-error solution requires that one of the two sample points have minimum radius for other types of curves. Various types of coupling functions have been examined numerically to determine if the result can be generalized and if there are counterexamples. The solution has been examined for amplitude coupling in the form of a power law $a(\psi) = \alpha[\psi/(2\pi)]^r + \beta$, where $\alpha + \beta = 1$ and $0 \leq \psi \leq 2\pi$. The exponent r was varied between 0.2 and 3. Although this testing is not exhaustive, for each desired complex value the minimum-error encoding once again used the sample point at either $(\beta, 0)$ or $(1, 2\pi)$. For each of these coupling functions, the amplitude increases monotonically with phase.

Table 1. Encoding Errors for Minimum-Error, Compensated, Maximum-Error, and Phase-only Pseudorandom Encoding for Various $\mathbf{a}_c$

a_0	ψ_0/π	ν/π	$\epsilon_{\min}$	ϵ_c	$\epsilon_{\max}$	ϵ_{po}
0.80	1.75	0.35	0.09	0.24	0.24	0.36
0.68	0.05	0.10	0.06	0.06	0.18	0.54
0.67	1.60	0.46	0.19	0.36	0.37	0.55
0.60	0.75	0.32	0.10	0.11	0.15	0.64
0.40	0.00	0.62	0.12	0.37	0.70	0.84
0.05	0.75	0.95	0.38	0.46	0.73	1.00
0.00	—	1.00	3/8	3/8 to 3/4	3/4	1.00

In search of a counterexample, other coupling functions have been examined that are nonmonotonic. For simulations using these functions, neither of the two sample points is minimum in amplitude for minimum-error encoding. One of these coupling functions that provide a counterexample is $a(\psi) = 0.75 + 0.25 \cos \psi$. Another function is that for the phase-only characteristic. It has the property that every pair of sample points that are collinear with the desired complex value produce the identical amount of encoding error. The geometry is unchanged by a shift of the origin, so that the encoding error as calculated by Eq. (32) remains constant as long as the modulation characteristic is perfectly circular on the complex plane.

Summary. This section has described a method for encoding based on minimizing pseudorandom encoding error. The method has been specialized so that only binomial distributions are admitted. An even more involved problem would be the search for an arbitrary density function that minimizes the encoding error. This possibility has yet to be considered. Also, the higher-order moments (e.g., the variance of the intensity[6]) provide additional information on the noise and the errors in the diffraction pattern. This information could be used to decide between two encodings that generate the same amount of noise ϵ (as is the case for phase-only and other circular characteristics). Although much more analysis is possible, the analysis procedure presented here does permit comparisons of encoding error for the compensated algorithms with the minimum- and maximum-error limits. The compensation method, although it does produce more error, is more conducive to quick implementation. Furthermore, for modulation characteristics that are closer to circular, the amount of error produced by the amplitude compensation method would approach the minimum error even more closely.

8. DEMONSTRATIONS OF THE ENCODING ALGORITHMS

A better appreciation of the usefulness and the performance of the pseudorandom-encoding algorithms for coupled modulators can be gained by comparing their ability to encode a desired complex-valued function. This is done by way of computer simulation. The identical desired function is used for each encoding algorithm. The diffraction pattern of the desired function and of each encoded function is simulated and plotted on comparable scales. Three of the encodings use algorithms for coupled SLM's that are developed in this paper, and two encodings are previous methods that are included for comparison.

Procedure. For this study the SLM is assumed to be composed of 128 × 128 pixels. It is modeled as an array of 128 × 128 samples. The desired fully complex function is formed by adding together subarrays of sizes 128 × 32, 64 × 64, and 32 × 128. Each subarray has amplitude that is constant and phase that varies linearly with position. Thus the Fourier transform of each subarray would be a two-dimensional sinc function centered at a point that is determined by the slope of the phase. The phase slopes are chosen so that diffracted spots will be centered on the horizontal axis. The subarrays are positioned so that only two subarrays overlap at any position in the modulator plane. This gives the resulting amplitude modulation in Fig. 8(a) the appearance of vertical interference fringes. The fringe spacing varies, depending on which pair of subarrays overlap. It is desired that each spot have an identical peak intensity, and for this reason the magnitudes of each subarray are identical. This leads to modulation amplitudes that vary from 0 to γ. [Also note that for $\gamma = 1$ the diffraction efficiency, with the use of Eq. (6), is 0.5.]

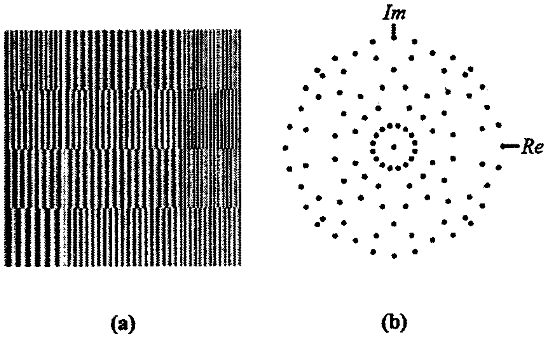


Fig. 8. Plots of the desired fully complex function $\mathbf{a}_c$ used in the simulation study: (a) desired magnitudes a_c as they would appear on the 128 × 128-pixel SLM, (b) desired values $\mathbf{a}_c$ shown on the complex plane. Each value shown occurs numerous times.

The resulting diffraction pattern is simulated by placing the 128 × 128 array of complex numbers in a 512 × 512 array of zeros and then performing the fast Fourier transform. A gray-scale rendition of intensity is shown in Fig. 9(a) for the central 65 × 512 samples of the fast Fourier transform, and an intensity plot is shown in Fig. 10(a) for the central 1 × 512 samples. The gray-scale rendition is saturated so that fully white corresponds to a level that is 15% of the peak intensity. Although the Fourier transform of each of the subarrays is a sinc function, some interference from sidelobes is also evident. Nonetheless, the eight spots are nearly identical in intensity.

Three pseudorandom-encoding algorithms for coupled SLM's are implemented by assuming the same two SLM coupling characteristics as those used in Sections 5–7. The first algorithm is the example 2 algorithm with use of the 4π coupling characteristic (the solid curves in Fig. 1). The magnitude of the desired function is scaled so that $\gamma = 0.25$. The second algorithm is the example 3 algorithm with use of the 2π coupling characteristic (the dotted curves in Fig. 1). The algorithm is augmented with method 3 to handle values that lie in the unrealizable

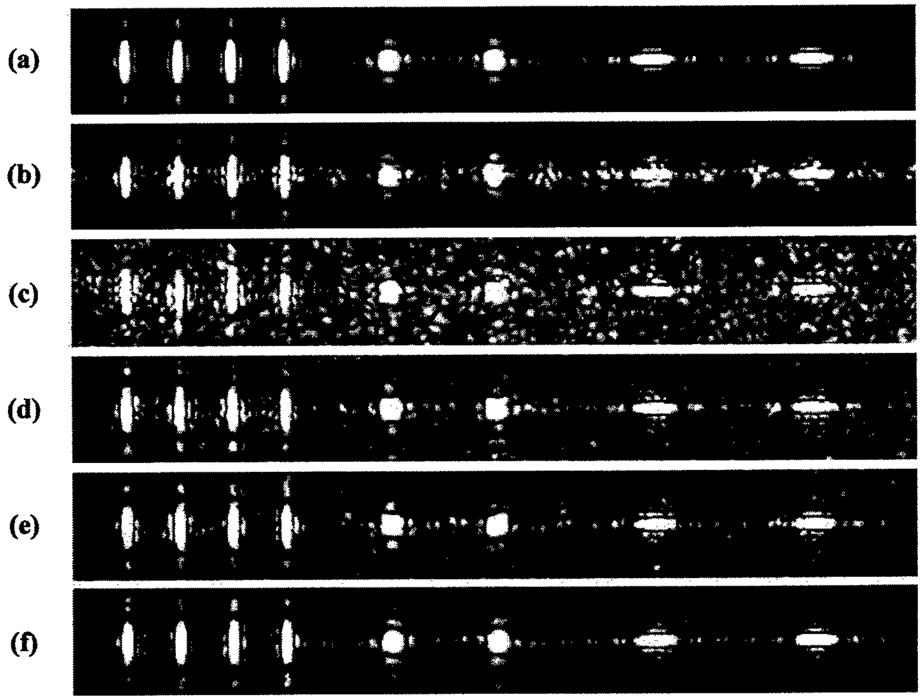


Fig. 9. Gray-scale plots of the intensity of the diffraction patterns resulting from (a) the desired fully complex function and (b)–(f) the various algorithms b–f. The algorithms are described in Section 8. Each image shows the central 65 × 512 samples of the simulated 512 × 512-sample diffraction pattern. In each image a fully white gray scale corresponds to an intensity that is 15% of the maximum spot intensity plus the minimum spot intensity divided by 2 [(max + min)/2].

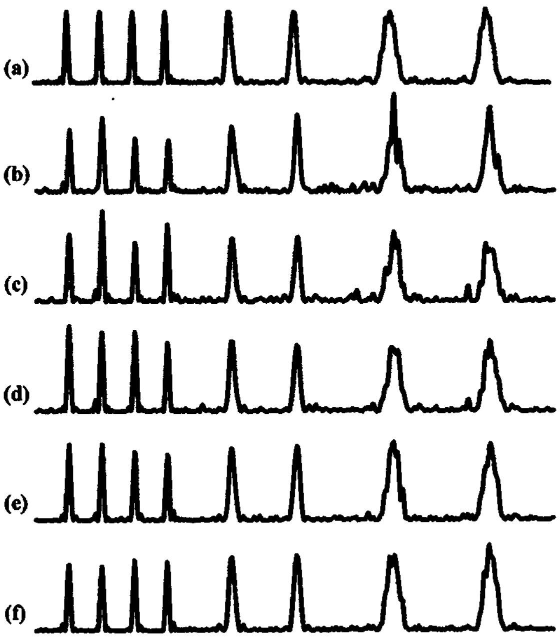


Fig. 10. Cross sections of the diffraction patterns resulting from (a) the desired function and (b)–(f) the various algorithms b–f. Each cross section is the central 1 × 512 samples of the simulated 512 × 512-sample diffraction pattern. Each trace is normalized so that the maximum spot intensity plus the minimum spot intensity divided by 2 [(max + min)/2] is of identical length on the vertical scale of each plot.

peninsular region of Fig. 5. The two possible values, a_l and a_u, that are used in Eq. (27) are chosen to be the upper and lower boundaries of the peninsula for a specified value of ψ_c. One or the other of these two boundary values (depending on the value of a random number and d) is then encoded by example 3. The desired function is scaled so that $\gamma = 0.55$. The third algorithm is the minimum-error encoding for binomially selecting one of two possible points. This algorithm is equivalent to method 4 with use of the minimum-error selection criteria given in Section 7. The amplitude coupling is also the 2π coupling characteristic (dotted curves of Fig. 1), and the scaling is $\gamma = 0.49$. For convenience I will refer to these three algorithms as c, d, and e, respectively, which are the same labels as those used for the corresponding results in Figs. 9 and 10.

There are two other algorithms that have been implemented for purposes of comparison. These are a deterministic algorithm that is somewhat similar to the phase-only filter, except that it is applied to the 2π characteristic of Fig. 1, and a pseudorandom encoding for a phase-only characteristic. These algorithms will be referred to as b and f. As with the phase-only algorithm, the nonrandom algorithm sets the actual phase ψ to the desired phase ψ_c. Thus the implemented modulation is $\mathbf{a} = a(\psi_c)\exp(j\psi_c)$. The pseudorandom phase-only encoding is identical to example 3 with the coupling characteristic set to $a(\psi) = 1$ for all values of ψ. The relationship between effective amplitude and spread [Eq. (24)] is plotted in Fig. 4 (thin solid curve). The spread is directly calculated as $\nu = \arccos(2a_c)$. For the phase-only algorithm the desired complex function is scaled so that $\gamma = 1$.

After each of the five encodings is performed, the data are zero padded and fast Fourier transformed by the identical procedure used for the desired fully complex function. Each gray-scale image in Fig. 9 is saturated to approximately 15% of its peak intensity. The maximum (max) and minimum (min) intensities of the eight peaks in the corresponding trace in Fig. 10 are found. Then the peak gray scale for the gray-scale image in Fig. 9 is set to 15% of (max + min)/2. Figure 10 is plotted so that the vertical scale is identical for the quantity (max + min)/2. The peak fluctuations are used to calculate a uniformity measure as well. This is $u =$ (max + min)/(2 × max).

A few comments on the numerical implementation of the algorithms may be of interest. The algorithms are all implemented by using MATHCAD whiteboarding software running on a 100-MHz Pentium personal computer with 32 Mbytes of memory and a Windows 95 operating system. The time required to encode the 128^2 desired complex values takes of the order of 2 s for algorithms b, c, and f and of the order of 3 min for algorithms d and e. No special effort has been made to optimize the implementations for computational speed or real-time implementation. Furthermore, since MATHCAD is an interpreter and a graphical interface, rather than a compiled language, the speeds reported here are not representative of what is possible with compiled code. To ensure consistency in comparing the various encodings, the identical 128^2 random numbers are used for each type of encoding. Algorithms b and f are directly implemented with standard functions, so no further discussion of their implementation is given.

Algorithm c requires that the amplitude of Eq. (21) be inverted to specify ν. This can be done by using a one-dimensional nonlinear equation solver or by developing a lookup table. Based on the smoothness of the curves in Fig. 3, the latter option has been pursued. A two-dimensional spline-fitting function is available in MATHCAD. It requires that the sample coordinates be rectangularly spaced in two dimensions. The way that this condition has been obtained is the following: (1) For a fixed value of desired phase ψ_c, calculate the magnitude of Eq. (21) divided by $a(\psi_c)$ for 11 values of ν between 0 and 2π. This produces values of $x = a_c/a(\psi_c)$ that are normalized between 0 and 1. (2) This calculation is performed for 11 values of ψ_c between $-\pi$ and π to yield 121 values. (3) The 11 pairs of values for each fixed value of ψ_c are fitted with a one-dimensional spline to produce 11 spline fits of ν as a function of x. (4) Each spline fit is then interpolated at identical values of x. This produces 121 values of ν on rectangular coordinates of x and ψ_c. (5) These values are then fitted by a two-dimensional spline. The appropriate value of ν is then found by supplying values of ψ_c and $x = a_c/a(\psi_c)$ to the MATHCAD two-dimensional spline interpolation routine. Such routines actually use a lookup table to localize the value of the function followed by interpolation to refine the accuracy of the value.

The development of a two-dimensional spline fit proved too difficult to apply to algorithm d. One problem is that desired/effective amplitude a_c is not always a monotonic function of ν (see Fig. 4). Thus it is not possible to exchange the ordinate and the abscissa, and then spline-fit ν against a_c over the full π range. A nonlinear equation solver has also been tried, but the possibility of two solutions for ν and the sharp discontinuity in the curves frequently lead to nonconvergence to a solution. A method that always seems to produce a solution is to limit the fit (and subsequently to interpolate) only over ranges for which a_c is a monotonic function of ν. These various ranges have been evaluated in Section 5 and are identified in Fig. 5. The description of the logical selection of the ranges is somewhat tedious. I will describe only one case for illustration. If ψ_c is a fixed value between $3\pi/2$ and 2π, a_c is a monotonic function for all values of ν between 0 and $2(2\pi - \psi_c)$. (The maximum value of a_c corresponds to the lower bound of the unrealizable peninsular region in Fig. 5.) A spline fit of ν as a function of a_c can be made over this range. Although it is numerically inefficient, the fit is repeated for each new value of $\mathbf{a}_c$ as follows: (1) For a given value of $\mathbf{a}_c$, the appropriate range of ν is identified; (2) 11 values of a_c are calculated over the range of ν for the given value ψ_c; (3) a spline fit of ν as a function of the a_c is performed; and (4) a spline interpolation is performed to calculate the required value of ν.

For algorithm e a nonlinear equation solver is employed to find the two sample points that pseudorandom-encode the desired complex value with a minimum error. The result from Section 7 is used that one of the two sample points of the minimum-error solution is usually (0.5, 0). By limiting the scale factor so that γ is less than 0.5 (as has been done), it becomes possible to avoid consideration of the other possibility [that one sample point can sometimes be (1, 0)] and always use (0.5, 0). Under this condition the equation solver converges for all 128^2 complex values. For γ of 0.5 or somewhat greater, the equation solver did not always converge. The solution is essentially identical to the result of the procedure given in Section 7. Given the points $\mathbf{a}_c$ and (0.5, 0), the solver finds the intersection point of the common line through these two points and the modulation characteristic.

Discussion of results. Figures 9 and 10 show that all five methods produce diffraction patterns that are similar to the desired diffraction pattern. Algorithm b, the deterministic algorithm, is the least uniform, having a uniformity of $u = 0.76$. The pseudorandom algorithms c, d, e, and f have greater uniformities, of $u = 0.82, 0.88, 0.92$, and 0.88, respectively. The uniformity of algorithm e may be overstated, since the peak intensity of its diffraction pattern (unlike the other curves) is on a line other than the one plotted in Fig. 10. If this value had been used instead, then u would be calculated to be 0.89. The key result to note is that the pseudorandom algorithms lead to more uniform or accurate reconstructions of the desired function than does the deterministic method. This correspondence between pseudorandom and nonrandom methods has been seen to a greater[10] or lesser[14] degree depending on the specific functions that are being encoded. A second observation is that, of the pseudorandom encodings c–e, c, which uses a low value of γ, is much less uniform than d and e, which use larger values of γ.

The algorithms are also compared on the basis of background noise. Algorithm b generates more intense noise than algorithms d and e and levels that are comparable with those of algorithm c. However, away from this axis the noise for algorithm b is substantially weaker than that for any of the other algorithms. More severe background noise has been noted in Ref. 10 for two-dimensional spot arrays produced by deterministic encoding. In the pseudorandom algorithms the background noise of algorithm c, as shown in Fig. 9, is much lower than that for algorithms d and e. Algorithm f has even lower levels of background noise. The background noise in algorithms c–f has the appearance of speckle, as expected. The higher levels of speckle noise in algorithm c are due to the small value of γ, which leads to a significant number of large phase spreads in the design algorithm.

The most pleasing result of this simulation study is that algorithms d and e, which are applied to SLM's with significant degrees of amplitude coupling, perform nearly as well as pseudorandom encoding (algorithm f) on phase-only SLM's. This result shows that pseudorandom encoding is of more than mathematical interest and that it has its uses in control of today's SLM's.

9. SUMMARY AND CONCLUDING REMARKS

The concept of extending pseudorandom encoding from phase-only to amplitude-coupled phase modulators has been explored. Including amplitude compensation in the probability density function has been used to derive solutions over a continuous range of effective amplitudes for a given value of effective phase. The encoding formulas found are only slightly more involved than those previously found for phase-only modulators. The example 3 algorithm, which uses the binomial distribution, is useful in that it is adaptable to a wide variety of amplitude-coupling characteristics. In developing an encoding algorithm, it is necessary to determine the range of values that can be encoded. Modifications (e.g., blending of multiple encoding algorithms) that extend this range can lead to improved performance. The maximum range that can be encoded by all possible pseudorandom algorithms has been identified as a convex region that is bounded (in part) by the modulation characteristic. This region has been identified by use of the properties of the binomial distribution. This analysis with the binomial has also been used to illustrate how various distributions can encode the same value, though with differing levels of error. The topics covered in this paper provide a framework for the development of pseudorandom-encoding algorithms for various coupled modulators.

Compared with previous encoding algorithms, pseudorandom encoding is novel in that it permits the direct encoding of complex-valued information to one pixel at a time with only a limited consideration of the settings of neighboring pixels. This has great utility for programming SLM's with fairly arbitrary complex-valued functions in real time. The idea that a limited set of modulation values can represent a continuum on the complex plane is in some sense parallel to Shannon's concepts on communication in the presence of noise.[15] In Shannon's theory the information content of a signal space is determined by the dimensionality of the signal space (i.e., the number of ways that a signal can be represented) divided by the volume occupied in this space by noise. In optical processing, the modulation characteristic of SLM's limits the area of the complex-valued space that can be addressed. With pseudorandom encoding, a continuum of the complex space can be addressed, but, as a result, white noise is generated. The proposed improvements involving the blending of various pseudorandom and other algorithms appear to be aimed at further increasing the available signal space.

ACKNOWLEDGMENTS

I thank Christy L. Lawson for assistance in preparing the simulations presented in this paper. This study was supported by Defense Advanced Research Projects Agency contract F19628-92-K0021 through Rome Laboratory, NASA cooperative agreement NCCW-60, and U.S. Office of Naval Research grant N00014-96-1-1296.

REFERENCES

1. R. D. Juday, "Correlation with a spatial light modulator having phase and amplitude cross coupling," Appl. Opt. **28**, 4865–4869 (1989).
2. C. Soutar and S. E. Monroe, Jr., "Selection of operating curves of twisted-nematic liquid crystal televisions," in *Advances in Optical Information Processing VI*, D. R. Pape, ed., Proc. SPIE **2240**, 280–291 (1994).
3. L. G. Neto, D. Roberge, and Y. Sheng, "Programmable optical phase-mostly holograms with coupled-mode modulation liquid crystal television," Appl. Opt. **34**, 1944–1950 (1995).
4. B. R. Brown and A. W. Lohmann, "Complex spatial filter," Appl. Opt. **5**, 967 (1966).
5. J. P. Kirk and A. L. Jones, "Phase-only complex-valued spatial filter," J. Opt. Soc. Am. **61**, 1023–1028 (1971).
6. R. W. Cohn and M. Liang, "Approximating fully complex spatial modulation with pseudo-random phase-only modulation," Appl. Opt. **33**, 4406–4415 (1994).
7. R. W. Cohn, A. A. Vasiliev, W. Liu, and D. L. Hill, "Fully complex diffractive optics by means of patterned diffuser arrays," J. Opt. Soc. Am. A **14**, 1110–1123 (1997).
8. D. C. Chu and J. W. Goodman, "Spectrum shaping with parity sequences," Appl. Opt. **11**, 1716–1724 (1972).
9. D. C. Chu and J. R. Fienup, "Recent approaches to computer-generated holograms," Opt. Eng. **13**, 189–195 (1974).
10. R. W. Cohn and W. Liu, "Pseudorandom encoding of fully complex modulation to bi-amplitude phase modulators," in *Diffractive Optics and Microoptics*, Vol. 5 of 1996 OSA Technical Digest Series (Optical Society of America, Washington, D.C., 1996), pp. 237–240.
11. R. W. Cohn and M. Liang, "Pseudorandom phase-only encoding of real-time spatial light modulators," Appl. Opt. **35**, 2488–2498 (1996).
12. A. Papoulis, *Probability, Random Variables, and Stochastic Process*, 3rd ed. (McGraw-Hill, New York, 1991), Chap. 5, pp. 101–102 and Chap. 8, pp. 226–229.
13. R. D. Juday, "Optimal realizable filters and the minimum Euclidean distance principle," Appl. Opt. **32**, 5100–5111 (1993).
14. L. G. Hassebrook, M. E. Lhamon, R. C. Daley, R. W. Cohn, and M. Liang, "Random phase encoding of composite fully-complex filters," Opt. Lett. **21**, 272–274 (1996).
15. C. E. Shannon, "Communication in the presence of noise," Proc. IRE **37**, 10–21 (1949).

threshold.[1–3] The realization of these properties depends, to a great extent, on the ability to fabricate quality surfaces for these crystals. The focus of this note is the precise hand polishing of these crystals .

Stain-free surfaces are the key to realizing the reported high-surface damage values of these materials.[1, 4] Reported here is a method for producing surfaces for BBO and LBO having surface figure of $^{1}/_{10}$th wave or better (at 633 nm), parallelism of less than 1 arc s, surface roughness of 10 Å RMS or less, and a starch and dig standard of at least 10–5 for crystals as large as 5 3 5 3 10 mm^3. These surfaces were generated using hand polishing as might be done for small quantities while conserving the scarce and expensive BBO and LBO materials. The following procedure given below takes between five and seven hours to complete.

Both BBO and LBO are hygroscopic and soft. Rubber surgical gloves are necessary when handling the crystals to prevent staining from hands. Potassium dihydrogen phosphate (KDP), is used as the blocking material because it is relatively inexpensive and has about the same hygroscopic susceptibility and softness as BBO and LBO. Rectangles of KDP are prepared to surround the BBO or LBO crystal. By having the KDP pieces all the same thickness, only slightly thicker than the BBO and LBO, grinding time can be saved. This thickness condition eliminates grinding the BBO and LBO and minimizes possible subsurface damage in these crystals. The resulting configuration is a square BBO or LBO crystal bordered by the KDP pieces. The KDP-BBO or LBO square is about 1.5 in. on a side. A low melting wax such as blanchard wax is used to glue these crystals to a 2" diameter pyrex or glass work piece. It is important that this gluing step be done in an oven rather than on a hot plate to avoid air currents that could cause crystal cracking.

The lapping is done using WCA-9T corundum grinding compound lubricated with ethylene glycol. Avoiding larger size grit minimizes subsurface damage. Grinding compound finer than WCA-9T appears to cause minimal removal while producing a large number of scratches. Grinding is followed by rough polishing using 3-µm diamond dust in silicone oil on a Pellon pad lap. For this step, all the gray is removed and the surface is 1 or 2 rings convex (at 633 nm). The final polish is done using a convex pressed pitch polisher (73 Gugolz, 6" diameter) and approximately 6 mg of $^1/_2$-µm diamond dust using ethylene glycol as the lubricant. For each operation a slowly rotating lap (5–10 rpm) is used while holding the work piece by hand.

The key step is the protection of the finished first surface during the removal of the BBO and LBO crystals from the work piece, and while polishing the second surface. The blanchard wax is dissolved using trichloroethylene (TCE) in a warm oven. Wax residue is removed by wiping with a lens tissue or Q-Tip and TCE. Surface degradation will occur with extended wiping or allowing the solvent to dry on the surface. Should this happen hand-held touch up is required. We do not know of any adhesive that will not stain the polished surface of KDP, BBO, or LBO.

To polish the second surface, while preserving the finished first surface, requires hand-work on the unmounted BBO or LBO sample, again emphasizing rubber gloves to prevent surface degradation. By hand, the second surface is ground and polished the same as the first. Parallelism is the added requirement for this step. Parallelism is verified first using an autocollimator and then an interferometer as the surfaces approach parallelism. Sometimes it is necessary to touch up the first surface as the final step.

The procedure given here results in negligible subsurface damage as indicated by the high value of threshold for laser damage. It is believed that the extensive polishing and minimal grinding contributes little to subsurface damage that may be the seed for the initiation of laser damage. Working in a "clean room" or glove box atmosphere with reduced humidity can help to reduce the water vapor produced stains. These controlled conditions were not available at the time.

References

1. C. Chen *et al.*, "New non-linear crystal: LiB_3O_5," J. Opt. Soc. Am. B **6**, 616 (1989).
2. S. Lin *et al.*, "The nonlinear optical characteristics of a LiB_3O_5 crystal," J. Appl. Phys. **67**, 634 (1990).
3. D. Eimerl *et al.*, "Optical, mechanical, and thermal properties of barium borate," J. Appl. Phys. **62**, 1968 (1987).
4. B.S. Hudson, "β-Barium borate: An important new capability in laser technology," Spectroscopy **2** (6), 33 (1987).

Frequency Swept Measurements of Coherent Diffraction Patterns

Markus Duelli, David L. Hill, and Robert W. Cohn, The ElectroOptics Research Institute, Univ. of Louisville, Louisville, Ky.

Abstract

Interference fringes arising from multiple reflections can significantly alter the diffraction patterns of diffractive optical elements. One way to reduce interference effects is by time-integrating the diffraction pattern while frequency sweeping the laser source. This method is especially useful when it is not possible to remove the cover glass from the observation camera.

The use of charge coupled device (CCD) cameras in optical systems, together with a laser light source, is widely applied to a variety of measurements. But coherent imaging can introduce severe alterations of the detected signal due to the unwanted interference of multiple reflections of the beam. These reflections that arise from various optical surfaces in the system, including the cover glass of the CCD chip, are difficult to completely eliminate. A classic solution is to use a spatially coherent broadband source.[1] An alternate approach, described in Reference 1, adds together a set of images, each formed with a different wavelength of spatially coherent narrowband light. While the emphasis of the earlier work was to reduce speckle, the procedure evidently reduces interference fringes as well (see Fig. 21.15 in Ref. 1). Today, with the availability of tunable laser diodes and CCD cameras, it appears possible to perform wavelength averaging in real-time. We will demonstrate this technique and report on the improvement in the accuracy and repeatability of the diffraction patterns produced by a set of diffractive optical elements (DOEs).

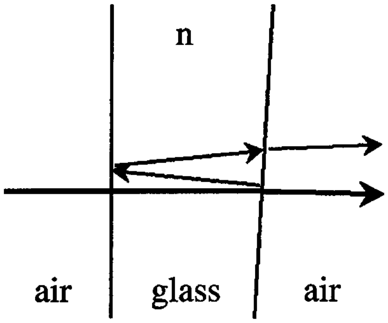


Figure 1. Multiple reflection between two interfaces.

Figure 1 illustrates a typical source of interference from Fresnel reflections in a glass plate. A fringe pattern is usually observed across the plate due to (even a slight) lack of parallelism between the two surfaces. The fringe pattern can be averaged out by continuously varying, by at least 2π the phase difference between the transmitted beam and the doubly reflected beam, and integrating the intensity pattern during the sweep time. A 2π phase change is achieved with a sweep range of

$$\Delta\lambda = \lambda^2/2nd = (\lambda/\nu)\,(c/2nd) \qquad (1)$$

where λ is the source wavelength, d is the separation between the two reflecting surfaces, and n is the refractive index (which is assumed to be constant with wavelength). A wavelength change of $\Delta\lambda = 0.25$ nm will produce a 2π shift for $\lambda = 860$ nm, $n = 1.5$, and a thickness $d = 1$ mm, a typical thickness for cover glass and planar DOEs. The second equality is written in terms of the source frequency ν and the speed of light c. Writing equation 2 this way identifies the frequency change $\Delta\nu = c/2nd$ *as the free spectral range of a Fabry Perot etalon.*[2]

In a preliminary experiment, a diode laser of nominal wavelength $\lambda = 860$ nm is used. A collimated beam is passed through a 3-mm microscope slide and the 1-mm cover glass of the observation camera (a cooled CCD camera with variable time integration) and is recorded by the camera. The observed intensity distribution is shown in Figure 2a. The larger period fringes are from the microscope slide and the smaller period fringes are from the cover glass. The temperature of the laser head, and thus the frequency of the emitted light, can be controlled by an external voltage. An input voltage between 1–4 V varies the temperature between 10–40°C. By supplying a time varying voltage the temperature changes accordingly. A low frequency step function (period $T = 60$ s) as the control voltage is used. This results in a temperature change that varies linearly with time between 24°–37°. As the temperature changes the fringe pattern is observed to translate. By exposing the CCD during one period of the fringe translation the interference pattern is averaged out. In our set-up one period of fringe translation, corresponding to the desired 0.25-nm wavelength change, occurres in 3 s. The pattern resulting after a 3 s exposure is shown in Figure 2b. While the fringes are averaged out, the frequency sweep has no visible influence on the diffraction pattern from the dust particles on the glass plate.

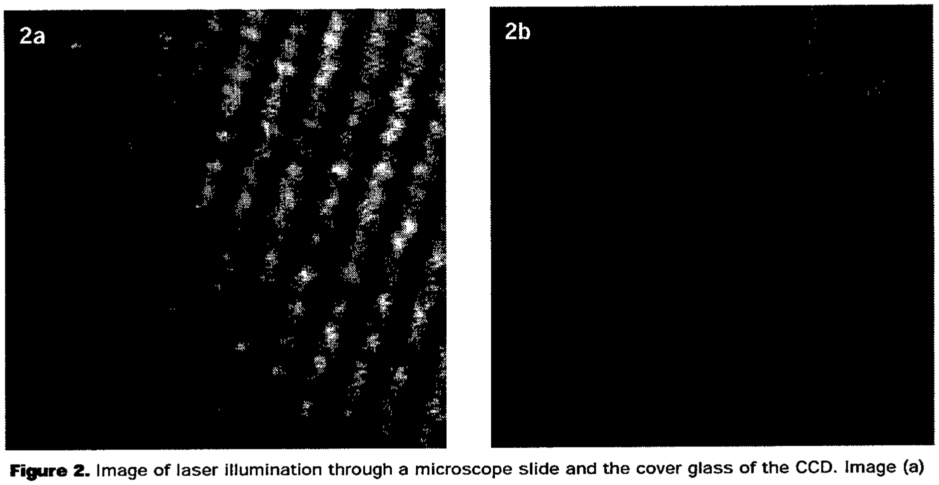


Figure 2. Image of laser illumination through a microscope slide and the cover glass of the CCD. Image (a) without and (b) with frequency sweep.

The swept frequency method is used to characterize a set of identically designed diffractive optical elements. These devices, when illuminated with collimated light, are designed to produce 64 spots of nearly equal intensity in the Fourier plane. The uniformity (defined as the standard deviation of the intensity of the spots) of the designed spot array, is calculated to be 7%. The DOEs are 300 3 300 pixel phase elements with each pixel set to one of eight possible phase levels. Four of the seven DOEs are anti-reflection coated on the backside of the glass substrate. In our measurements the DOE is illuminated with a collimated beam and the diffracted light is focused with a lens onto the CCD camera. The diffraction pattern is recorded on the CCD camera with and without frequency sweeping. With no frequency sweeping an average uniformity of 12.1% with a standard deviation of 1.5% is measured. There is no appreciable difference between measurements of antireflection and non-antireflection coated devices. This indicates that the disturbing reflections mainly originate from the cover glass of the CCD camera. With frequency sweeping the average measured uniformity of the seven devices is reduced to 7.9% with a standard deviation of 0.8%. The swept frequency method improves the repeatability of the uniformity measurement. In addition, the results compare more favorably with the theoretical levels. Other measurements including signal-to-peak background ratio and diffraction efficiency compare well with theory, though these measurements are not as sensitive to reflections as is uniformity.

The method is valid as long as the sweep range $\Delta\lambda$ does not introduce severe wavelength dispersion of the diffraction pattern. This is true as long as

$$\Delta\lambda/\lambda << w/f \qquad (2)$$

where f is the highest spatial frequency of interest in the diffraction pattern and w is the resolvable width of the optical system. For example, the diffraction pattern of an optical system that has a circular aperture is an Airy pattern of angular width $\theta_w = \tan^{-1}(w/z) \approx 1.22\lambda/l$ where l is the diameter of the aperture and z is the distance from the aperture. For the measured diffractive optic reported above the aperture is 4 mm and $\theta_w = 0.015°$. Furthermore, through numerical integration of the Airy intensity pattern over the sweep range $\Delta\lambda = 0.25$ nm the wavelength dispersion introduces less than 5% reduction in intensity for angles less than 26.9°. In the measurements reported, dispersion is negligible since the highest frequency spot from the array generator is at an angle of 0.85°.

In conclusion, the method of time averaging while frequency sweeping the source can be used to suppress interference fringes arising from multiple reflections. This method can more accurately measure the performance of diffractive optical elements. It is worth noting that by sweeping a laser in under 1/30 s it would be possible to eliminate multiple reflections as observed on a live video camera. This speed was not possible with temperature tuning, nor was it possible to tune the laser over an adequately large range by adjusting the laser current. However, we have reviewed the specifications for several commercially available, external cavity laser diodes, and find that with the fastest sweep rates (6–10 nm/s) fringe suppression can be observed with a live video camera.

Acknowledgments

The diffractive optical elements were fabricated by Honeywell as part of the DARPA diffractive optical element foundry run. This study was sponsored by NASA cooperative agreement NCC5-222 and ONR grant N00014-96-1-1296.

References

1. G.O. Reynolds *et al.*, *Physical Optics Notebook: Tutorials in Fourier Optics* (SPIE Optical Engineering Press, Bellingham, Wash., 1989), pp. 204–219.
2. G.O. Reynolds *et al.*, *Physical Optics Notebook: Tutorials in Fourier Optics* (SPIE Optical Engineering Press, Bellingham, Wash., 1989), pp. 278.

High Resolution Moiré Photography: Extension to Variable Sensitivity Displacement Measurement and to the Determination of Direct Strains

Pramod K. Rastogi, Swiss Federal Institute of Technology, Stress Analysis Laboratory, Lausanne, Switzerland.

Abstract

A method for obtaining a direct full field display of in-plane strain contours is demonstrated. On another front, the paper proposes the basis of a multi-sensitivity high resolution moiré photography system for in-plane displacement measurement.

High resolution moiré photography[1-3] is an important technique for the measurement of in-plane displacements of deformed objects. The method has many desirable features such as white-light object illumination, low sensitivity, whole field mapping of in-plane displacements, and the ability to be applied to specimens of largely varying sizes. The method uses the unique imaging properties of a lens covered with a mask containing two parallel slits. A so-masked lens system is used to image a periodic pattern, fixed on the object surface, onto a high resolution photographic film. This type of system has two main advantages. First, it serves to enhance the resolution over a narrow band of spatial frequencies in a direction parallel to the line joining the center of the two slits. Second, it increases the depth of focus by a significant amount.

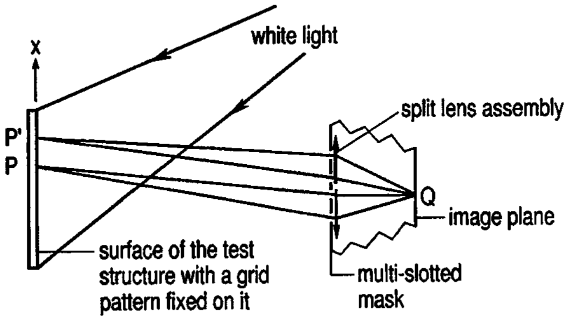


Figure 1. Schematic of high resolution moiré shearography for obtaining in-plane strain contours.

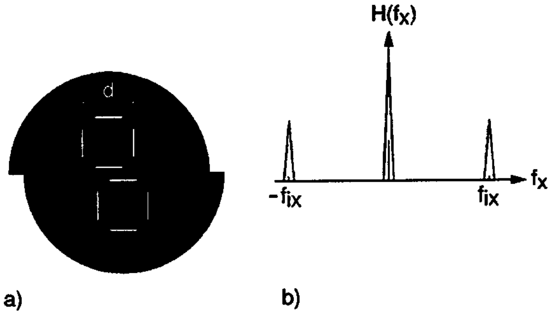


Figure 2. (a) Schematic of the two four-slot pupil arrangements covering the split-lens device and (b) its corresponding monochromatic optical transfer function along the x direction; b shows 2x magnified view.

In this paper, a novel scheme that combines the principles of optical shearing and a slotted mask arrangement to obtain directly the whole-field mapping of in-plane strains of a deformed object is introduced. The second aim of the paper is to extend the method's capability to include multiple frequency channels. The multi-frequency transmission is made possible by the use of a novel aperture masking arrangement, which during reconstruction permits observation in all bands. An immediate fall out of this proposal is the development of a multi-sensitivity high resolution moiré photography system for displacement measurement.

Strain measurement

The specimen under investigation is imaged by an image-shearing device on to high-resolution photographic film (see Fig. 1). The image-shearing device consists of a split lens assembly with each half of the assembly covered by a mask

Fully complex diffractive optics by means of patterned diffuser arrays: encoding concept and implications for fabrication

Robert W. Cohn, Anatoly A. Vasiliev, Wenyao Liu,* and David L. Hill

Department of Electrical Engineering, University of Louisville, Louisville, Kentucky 40292

Received February 13, 1996; revised manuscript received July 10, 1996; accepted December 11, 1996

Arbitrary complex-valued functions can be implemented as arrays of individually specified diffusers. For any diffuser, only average step height and vertical roughness are needed to control phase and amplitude. This modulation concept suggests potentially low-cost fabrication methods in which desired topographies are patterned by exposing photoresist with partially developed speckle patterns. Analyses and experimental demonstrations that illustrate the modulation concept and aspects of the fabrication method are presented, with particular emphasis on limitations of complex recording set by various photoresist and exposure properties. Applications of diffuser array concepts to spatial light modulators and to gray-scale lithographic printing of micro-optics are also mentioned. © 1997 Optical Society of America [S0740-3232(97)02305-3]

Key words: signal synthesis, diffractive optics, computer-generated holography, laser speckle, diffusers, phased arrays, gray-scale lithography, statistical optics.

1. INTRODUCTION

The properties of random phase have been widely applied to analyze the scattering of monochromatic light from rough surfaces.[1,2] The inverse problem of specifying the statistical properties of phase-only structures so as to obtain desired far-field diffraction patterns has received little attention. Recently, Cohn and Liang introduced a point-oriented encoding method, referred to as pseudorandom phase-only encoding, in which the phase modulation $\psi(x, y)$ is treated as a nonstationary process in the coordinates x and y.[3] Specifically, the statistics at any point of the modulation are selected so that the statistical average of the random phasor $\exp(j\psi)$ equals the desired fully complex modulation $a_c \exp(j\psi_c)$ at that same point. The far-field diffraction pattern from this random phase modulation approximates the desired diffraction pattern in the sense of the law of large numbers; that is, as a number of repeated statistical trials, N, is increased, the resulting diffraction pattern more accurately corresponds to its expected value or ensemble average.[4] The averaging mechanism here is the superposition in the far field of wave fronts that originate from points across the modulation.

To our knowledge, there is no prior art in computer-generated holography in which statistical properties have been varied with position to achieve design objectives. There are many applications in which pseudorandom phase codes are widely used in optical information processing, holography, and optical memory storage; however, these approaches all appear to use pseudorandom phase sequences that are stationary, as opposed to our approach, in which the statistics are nonstationary. Somewhat similar to pseudorandom encoding is the Davis–Cottrell method of randomly multiplexing two (or more) phase-only functions.[5] The two individual modulations are randomly interleaved. The probability of selecting one or the other phase function determines the relative strengths of the individual functions and their diffraction patterns. The random selection is nonetheless a stationary process, and the method does not permit the encoding of arbitrary complex functions.

Research on pseudorandom encoding has so far been directed toward real-time programming of spatial light modulators (SLM's) for pattern recognition filters and toward laser-beam steering and shaping operations.[3,6,7] In this study we consider the application of the encoding concept to fixed-pattern diffractive optics. Mathematically, there is no difference between phase-only SLM's and phase-only diffractive optics, but there are practical differences that suggest entirely new realizations and methods of fabrication. These differences include the following:

- The time available to design and fabricate a diffractive optic is much longer than that for real-time programming of SLM's.
- The feature size of diffractive optics (micrometer scale) is usually much smaller than that for SLM pixels (10 to 200 μm), and the spatial bandwidth of diffractive optics is correspondingly larger than that for SLM's.

One implication of increased bandwidth (which is proportional to the number N of independently controllable phase-only pixels) is that the diffraction patterns from pseudorandom encoding will more closely approximate (in the sense of the law of large numbers) the desired diffraction pattern. But there are fabrication costs involved in using the highest resolution. Fine features are typically written point by point with direct-write laser-beam or electron-beam patterning systems. Scanning of laser beams is usually quite slow in order to minimize vibrations.[8] Electron-beam scanning can also be slow if there are a large number of mechanical steps between fields. Furthermore, electron-beams are generally expensive to purchase and maintain.

These issues have led us to consider a class of design problems in which the desired complex-valued modulation $\boldsymbol{a}_c \equiv a_c \exp(j\psi_c)$ has a bandwidth that is substantially less than the bandwidth (i.e., the reciprocal of resolution) of the diffractive optic element. When this is true, group-oriented encoding can be used in place of point-oriented encoding.[9,10] But, again, group-oriented encoding suggests the need for higher-resolution pattern generators. Perhaps, rather than to direct-write each resolvable point in sequence, a reasonable compromise for this class of diffractive optic design problems is to pattern the entire group in a single processing step. Optical patterning of groups is reasonable given that the system provides an adequate variety of patterns and that the complexity of configuring the patterns is not too great. Furthermore, to justify using such a system in place of current direct-write systems, increased speed and less critical optomechanical tolerances are necessary.

In this paper we examine a concept for effectively achieving complex modulation. It is a direct extension of pseudorandom encoding in that any one group is an array of random phase shifts all drawn from the same statistical distribution. We show that each group can be realized as a diffuser pixel having a specified roughness and average step height. We will refer to a diffractive optic composed of an array of custom diffuser pixels as a patterned diffuser array. We describe a fabrication procedure that appears capable of the simplicity, the robustness, and the speed needed to supplant electron-beam and laser writers for fabricating group-oriented designs. Theory, simulations, and demonstration experiments are presented to illustrate the modulation concept and to evaluate the practicality of the proposed patterning procedure. A major emphasis of the evaluation of the patterning procedure is the nonlinear transformation of the random statistics of the exposure pattern (typically, laser speckle patterns) into the desired complex modulation.

2. CONCEPT: EFFECTIVE COMPLEX MODULATION OF DIFFUSERS AND DIFFUSER ARRAYS

The complex-modulating property of diffusers can be appreciated qualitatively by considering the effect of the surface texture of a diffuser on its far-field diffraction pattern (Fig. 1). The diffraction pattern from a smooth surface is a specular intensity pattern. The pattern from a rough surface is a diffuse pattern of noise that is also referred to as a fully developed speckle pattern (only the envelope of diffuse scatter is illustrated in Fig. 1). The grain size of the roughness pattern is typically much smaller than the area of the diffuser that is illuminated; thus the envelope of the speckle pattern is typically much more broadly spread than the specular component. The far-field pattern from a surface of intermediate roughness will contain both specular and diffuse components and is referred to as a partially developed speckle pattern. Thus it is possible to use diffuser surface roughness to attenuate the specular component. The intensity transmittance could be varied from unity to practically zero. Since the roughness is of a much higher spatial frequency than that of the illumination footprint, the specular component can be much brighter than the diffuse component, even for large values of attenuation. This qualitative description shows the sense in which diffusers control the amplitude of the specular component. Rather than absorbing light with a true attenuator, the unwanted light is scattered into the diffuse component. Of course, the phase can be changed by translating the diffuser along the optical axis.

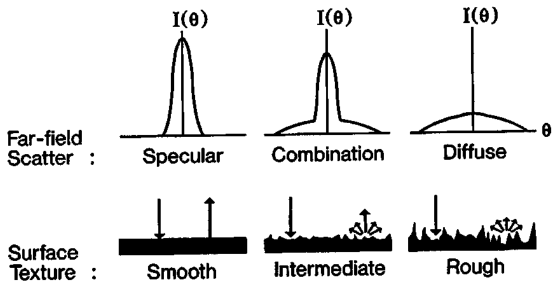


Fig. 1. Controlling specular intensity by varying surface roughness.

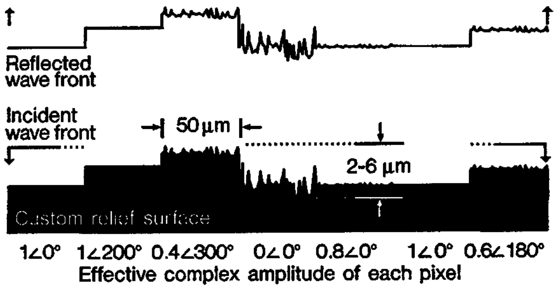


Fig. 2. Array of diffusers that produces a custom complex-valued modulation.

Our approach for encoding complex-valued designs is to use arrays of diffusers, with each diffuser representing a custom value of modulation (as illustrated in Fig. 2). Nearly arbitrary diffraction patterns, limited only by speckle noise, can be produced by superposition in the far field of the specular components from the individual diffuser pixels. A mathematical description of the complex-modulating property of diffusers can be explained in terms of the pseudorandom-encoding algorithm,[3] which is reviewed in general terms in this section and then specialized in Section 4 to accommodate the physical constraints set by the fabrication process. This section also compares, by way of example designs, the improvement made possible by using diffuser arrays in place of individual phase-only pixels.

A. Mathematical Description of Effective Complex Modulation and Pseudorandom Encoding

The modulation of a plane wave reflected from a phase-only surface is represented by the complex-valued (indicated by boldface type) function $\boldsymbol{a}(x, y) = \exp[j\psi(x, y)]$. The far-field diffraction pattern of the modulation pattern is $\boldsymbol{A}(f_x, f_y) = \mathscr{F}[\boldsymbol{a}]$, where $\mathscr{F}[\cdot]$ is the Fourier transform operator. Since the Fourier transform and the ensemble average are both linear, the average complex-valued far-field pattern of a random modulation can be written as

$$\langle \boldsymbol{A} \rangle = \mathscr{F}[\langle \boldsymbol{a} \rangle], \tag{1}$$

where $\langle \cdot \rangle$ is the expectation operator. Under the assumption that the random samples of $\boldsymbol{a}$ are statistically independent of position, the expectation of I, the far-field intensity pattern, is

$$\langle I \rangle = \langle |\boldsymbol{A}|^2 \rangle = |\langle \boldsymbol{A} \rangle|^2 + \langle I_s \rangle, \tag{2}$$

where $I_s(f_x, f_y)$ is a residual noise pattern that is due to the random phasings in the far field.[3] As long as the noise [represented by the second term of Eq. (2)] is adequately low, then Eq. (2) is approximately the magnitude squared of Eq. (1). Thus the specular and diffuse components of arbitrary diffraction patterns are identified with the two terms in Eq. (2). In the average sense of Eq. (2), any complex-valued modulation can be represented by the random phase-only modulation $\boldsymbol{a}(x, y) = \exp[j\psi(x, y)]$ by using the relationship

$$\langle \boldsymbol{a} \rangle = \int p(\psi)\exp(j\psi)\mathrm{d}\psi = a_p \exp(j\phi_p), \tag{3}$$

where $p(\psi)$ is the probability density function (pdf) of the phase and a_p is the resulting expected amplitude modulation. We will often refer to a_p as the effective amplitude, ϕ_p as the effective phase, and $\boldsymbol{a}_p \equiv (a_p, \phi_p) \equiv \langle \boldsymbol{a} \rangle$ as the effective complex amplitude or modulation.

The desired modulation $\boldsymbol{a}_c(x, y)$ is pseudorandom encoded by specifying a pdf $p(\psi)$ in Eq. (3) that gives $\boldsymbol{a}_p = \boldsymbol{a}_c$. The actual value of phase is selected by using a pseudorandom-number generator that has the required density function. Amplitudes can be encoded by using any number of pdf's in Eq. (3). A particularly useful family of pdf's is the uniform family of density functions, with spreads $\nu \in [0, 2\pi]$ and phase bias $\phi_p = \psi_c$. These densities, when evaluated in Eq. (3), give all values of amplitude between 0 and 1, according to

$$a_p = \mathrm{sinc}(\nu/2\pi). \tag{4}$$

Thus the correct density function for encoding a particular value of $a_p = a_c$ is found by inverting Eq. (4) for the appropriate value of ν. The most widely available random-number-generator routine is uniform with a spread of 1 and a mean of 1/2. A number selected by this routine would be scaled by ν and offset by $\phi_p - \nu/2$ to produce the actual random phase ψ. This procedure is applied at every coordinate point to calculate the functions $\nu(x, y)$ and $\phi_p(x, y)$, which in turn are used to determine the analog phase-only function $\boldsymbol{a}(x, y) = \exp[j\psi(x, y)]$ that represents the desired modulation $\boldsymbol{a}_c(x, y)$. Since most of our designs and analyses use discretely sampled functions, we often find it convenient to represent our functions as an array of N samples indexed in i (for example, ν_i and $\boldsymbol{a}_i$).

Other pdf's than the uniform can be used for $p(\psi)$ in Eq. (3). In Section 4 our fabrication approaches drive us to consider exponentially distributed (and other) phase statistics. These statistics raise additional challenges in that the phase modulation range of the diffractive optic can be many times larger than 2π for small values of effective amplitude (say, $a_p = 0.01$ or less).

B. Advantage of Diffuser Pixels over Single-Step Pixels: Directionality Gain

Equation (3) describes the effective complex modulation of a phase-only pixel, a diffuser, or arrays of either. Diffusers are typically modeled as arrays of random phases all drawn from the same random distribution. The only mathematical difference between a single pseudorandom phase-only pixel and a diffuser pixel is that the diffuser represents repeated statistical trials of the single pixel. According to the law of large numbers,[4] increasing the number of random trials associated with the diffuser pixel will make the far-field diffraction pattern more predictable; i.e., the specular component will be more clearly seen over the noise for diffusers having a larger number of phase samples.

This effect can also be interpreted as a directionality gain of the specular component over the diffuse component. If there are N statistically independent roughness samples, or cells, filling an aperture, then the intensity of the diffuse pattern will be reduced by a factor of $1/N$ over that resulting from one roughness cell filling the aperture. Since the diffraction pattern of the single roughness cell is identical to the pattern of the uniformly illuminated aperture, then the directionality gain of specular to diffuse is N.

To demonstrate more clearly the improvement possible by using diffusers in place of single-phase pixels, we compare the performance of encoding a specific complex-valued function into single-phase pixels and into diffuser pixels. The Fourier transform of the complex function produces an 8 × 8 array of equal-intensity spots. A single-phase pixel is modeled as a 3 × 3 array of identical phases, and a diffuser pixel is modeled as a 3 × 3 array of nonidentical phases. Both structures, each an array of 100 × 100 pixels, are pseudorandom encoded by using Eq. (4) to determine the spread ν_i of the random distribution associated with the ith pixel. The only difference is that for a diffuser pixel there are now nine phases, instead of one phase, selected from the uniform random distribution having spread ν_i and phase bias $\phi_{p,i}$. Another way to describe the difference between the two modulations is that the diffuser array consists of nine spatially multiplexed single-phase pixel encodings, each encoding performed by using a different random seed.

The resulting diffracted intensity patterns are shown in Fig. 3. The top row of Fig. 3 shows the diffraction patterns as simulated by using the fast Fourier transform, and the bottom row shows the result for diffraction from a Hughes birefringent liquid-crystal light valve that is programmed to approximate the desired phase modulation. As anticipated, the photographs show that the speckle is more broadly scattered and its intensity is reduced by using diffuser arrays.

Numerical measures of the quality of these diffractions patterns are presented in Table 1. The nonuniformity is defined as the standard deviation of the intensity of the 64 spots divided by their average intensity. The signal-to-noise ratio is defined as the ratio of the average intensity of the 64 spots to the average background intensity. For the experimental measurements the background intensity is calculated only in the vicinity of the spot array, while for the simulated diffraction pattern the entire pat-

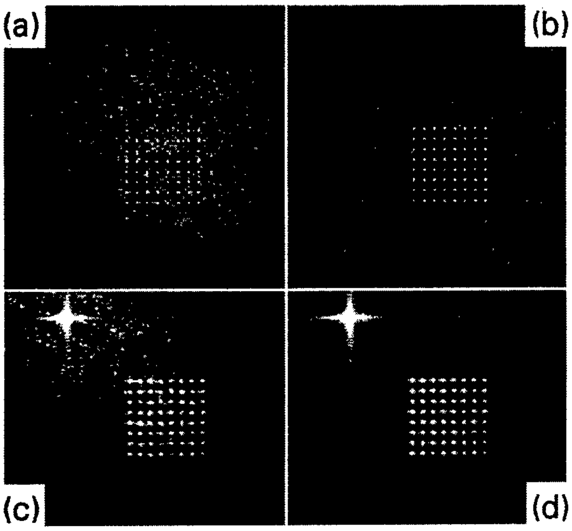


Fig. 3. Comparison of diffraction patterns from random encoding a 100 × 100 array of desired complex values to a 100 × 100 array of phase-only pixels and to a 100 × 100 array of diffuser pixels. Each diffuser pixel is a 3 × 3 array of phases that are randomly encoded to produce the same effective value of amplitude a_p. Shown are gray-scale images of diffraction pattern intensity for arrays of (a) phase-only pixels, theory; (b) diffuser pixels, theory; (c) phase-only pixels, experiment; and (d) diffuser pixels, experiment. The on-axis or dc component [upper left of (c) and (d)] is primarily due to Fresnel reflection from the cover glass, which has not been antireflection coated for this liquid-crystal light valve.

Table 1. Measures of Improvement of the Spot Array Design by Using Diffuser Pixels in Place of Single Phase-Only Pixels

	Nonuniformity (%)		Signal-to-Noise Ratio	
Diffuser Resolution	3 × 3	1 × 1	3 × 3	1 × 1
Simulation	9.9	23.6	1639	72
Experiment	21.6	29.4	35	14

tern is used. For each measure and for both simulation and experiment, the diffuser array has noticeably better performance.

For the experimental intensity patterns, the nonuniformity for the single-phase and diffuser pixel devices was originally measured as 36.9% and 30.4%. However, the images indicate that the intensity gradually decreases with distance from the optical axis. This is due to limited resolution of the SLM (which includes rolloff in the video output of the frame grabber and the cathode-ray tube that is the write light source for the light valve). The values in Table 1 for this experiment report nonuniformity with the linear and quadratic trends in both x and y removed (through the application of a least-squares regression). As further evidence that the rolloff is systematic, the nonuniformity for the theoretical images can be reduced only approximately 2% further (than reported in Table 1) by removing linear and quadratic trends. Even though the experimental spot arrays are less uniform and noisier than theory (because of loss of resolution and inexact phase control of the SLM), the improvements made possible by using diffuser pixels are apparent. Furthermore, applying pseudorandom encoding to nonideal devices produces results that are qualitatively similar to theory.

As an additional point of comparison, we have also calculated the performance of encoding the desired complex function $\mathbf{a}_a(x, y)$ to 300 × 300 single-phase pixels. The complex function is sampled with a higher resolution to 300 × 300 points instead of 100 × 100 points. The nonuniformity, found by simulation, for this single-phase pixel device is 6.8% as compared with 9.9% for the diffuser array. It is not surprising that the performance of the diffuser array is somewhat less than that of the 300 × 300 pixel modulator. However, there is a practical advantage to the diffuser in terms of simplicity of fabrication (described in Section 3). Furthermore, much improved performance is possible by further increasing the resolution of the diffusers.

C. Relationship to Prior Kinoform Design Procedures

Currently, numerically intensive global search and optimization algorithms are widely used for synthesizing modulation functions under the constraint of phase-only (in many cases binary phase-only) modulation.[11–15] Direct pixel-by-pixel or point-by-point encoding can be a practical alternative. Several methods of encoding complex functions onto phase-only diffractive structures were developed shortly after the introduction of the kinoform.[9,10] The most direct is the Kirk–Jones method,[16] in which a periodic carrier of spatial frequency f_0 that is modulated in amplitude α and phase ψ_a is converted into the phase-only function

$$\begin{aligned}\mathbf{a}(x, y) &= \exp[j\psi(x, y)] \\ &= \exp\{j[\alpha h(2\pi f_0 x) + \psi_a(x, y)]\}. \qquad (5)\end{aligned}$$

One specific case considered by Kirk and Jones was for $h(\cdot) = \cos(\cdot)$. For this case the Fourier-series expansion of Eq. (5) produces a dc component of complex amplitude

$$\mathbf{a}_c \equiv a_c \exp(j\psi_c) = J_0(\alpha)\exp(j\psi_a), \qquad (6)$$

where $J_0(\alpha)$ is the zero-order Bessel function. Thus $\mathbf{a}_c$ is proportional to the complex amplitude of the dc or zero-order far-field diffraction pattern. Any desired value of amplitude a_c between 1 and 0 can be implemented by inverting $J_0(\alpha)$ to find the appropriate value of α. Similar results can be developed for $h(\cdot)$ a square-wave carrier and also for a rectangular carrier of variable duty cycle.

From the perspective of the Kirk–Jones approach, patterned diffuser arrays use a random carrier. That is to say, rather than use a single-frequency carrier $h(\cdot)$, one adopts a carrier that is a randomly phased combination of a continuous range of frequencies. Whereas the traditional Kirk–Jones method scatters unwanted energy into the off-axis harmonics at discrete frequencies, a random carrier scatters unwanted energy uniformly (on average) into a continuous range of frequencies. For the single-frequency carrier approach, the unwanted harmonics are spatially separated from the desired signal. For the random-carrier approach, the noise and the signal occupy the same space. However, since the noise is spread uniformly over the entire observation space, the noise energy is often low enough to ignore.

For any of these carrier-based methods, it is important to note that the maximum useful diffraction efficiency of $\boldsymbol{a}(x, y)$ is limited only by the efficiency of the desired complex modulation $\boldsymbol{a}_c(x, y)$. Thus there is no implementation loss for the on-axis diffraction order. Furthermore, the optimization of the function $\boldsymbol{a}_c$ required to meet a specific set of design criteria is decoupled from the constraints imposed by the phase-only implementation. This could potentially lead to simplified and improved diffractive optic design procedures. (For instance, noniterative optimal window design procedures become possible; see Ref. 3 for a specific design of a top-hat far-field pattern.)

3. MICROTOPOGRAPHIC PATTERNING METHODS

A. Comparison with Prior Fabrication Methods

Kirk and Jones[16] also presented a fabrication procedure in which a photomask having a sinusoidally varying intensity transmittance is placed in contact with a photographic recording medium for which thickness depends linearly on exposure energy. The medium is exposed with an intensity pattern proportional to the function $\alpha(x, y)$. Then the mask is removed, and the medium is further exposed with a second pattern proportional to $\psi_a(x, y) = \psi_c(x, y) + 2\pi - \alpha(x, y)$ that adjusts thickness to produce the desired phase modulation $\psi_c(x, y)$. [The term $2\pi - \alpha$ compensates for the average thickness variations introduced by $\alpha h(\,\cdot\,)$.]

If a square-wave carrier is used instead of a sinusoid, the photomask is much easier to produce. If a rectangular carrier is used, the duty cycle is varied. This has the advantage that every pixel can be exposed with the same dose, but it has the disadvantages that the photomask must be written with extreme precision and a custom photomask is needed for each new device design. Also, all three deterministic carriers (sinusoidal, square, and rectangular) require two exposures to produce a desired complex value at a point. A single-exposure method can be envisioned in which laser interference is used to produce sinusoidal fringes and beam balance is adjusted to control phase bias. This pattern would be projected through a small aperture, and the entire photographic medium would be exposed by translating it under the aperture. This method, of course, requires good fringe stability.

The Kirk–Jones approach does not seem to have been widely used, apparently because of the requirement for analog control of the exposure. Currently, it is most common to fabricate computer-generated diffractive optic elements as binary and m-ary phase steps. However, lately there has been considerable progress in producing analog phase-only relief structures. Various approaches include projection printing and laser-beam or electron-beam direct write onto photoresist.[8,17–20] While the current direct-write systems accurately and precisely write topographic patterns into resist, they also are slow and expensive. As a result of the increasing emphasis on and success of custom-fabricated diffractive optics, we propose an alternative patterning approach; specifically, we consider the possibility of producing patterned diffuser arrays and the technical issues that would affect the quality of the resulting diffraction patterns.

B. Proposed Patterning Method

Our goal is to develop a robust, repeatable, and easy-to-implement patterning technique. While, in concept, we can write one random phase at a time by direct pseudorandom encoding [Eqs. (3) and (4)], there is really no need for this precise and detailed control. Instead, we can directly use the statistical properties of laser speckle, which are known to be reproducible and controllable.

Figure 4(a) illustrates one basic pattern generator concept. This apparatus is a type of proximity printer. An aperture (perhaps patterned on a chrome photomask) having the area of a diffuser pixel is kinematically supported as close to the photoresist as practical. The photoresist is exposed through the aperture, and then the substrate is translated to the next location to be exposed. The high-spatial-frequency random carrier is a fully developed speckle pattern generated by the ground-glass diffuser. An average intensity offset needed to produce a phase bias can be generated by temporal averaging of speckle patterns. This can be achieved, as illustrated in Fig. 4(a), by spinning a ground-glass diffuser with a constant angular velocity. The radial separation between the beam and the diffuser axis determines the linear velocity of the diffuser. Linear velocity together with exposure time then determines the effective bias. A theory for this is described in Section 4.

Figure 4(b) shows a modified approach in which a uniform intensity pattern can also be used to provide a phase bias. Statistical properties of the intensities of coherently biased speckle patterns are described in Ref. 1. We specifically consider the case in which the bias and the speckle pattern are mutually incoherent. For the second

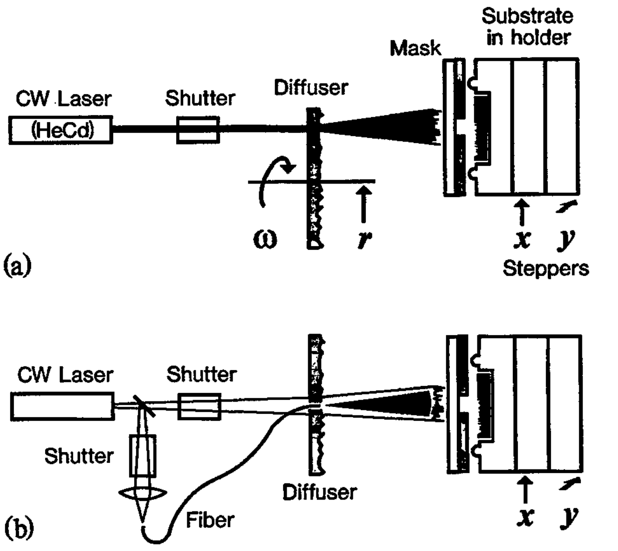


Fig. 4. Proximity exposure systems for producing complex-valued pixels. Phase offsets produced by (a) time-averaged recording of speckle patterns from a spinning diffuser and (b) adding a spatially uniform exposure, which, as shown, is derived from a single-mode optical fiber used as a point source.

approach the uniform and speckle illumination could obviously be combined with a beam splitter. However, in order to eliminate beam-splitter loss and multiple reflections, it is possible to bring a uniform coherent illumination through a small aperture (say a fiber optic) in the diffuser, as illustrated in Fig. 4(b). Mutual coherence between the spatially uniform and nonuniform sources can be achieved by rotating a polarized fiber into the cross-polarized state or by using the fiber to introduce a delay difference in excess of the coherence length of the laser.

A third approach would be simply to apply appropriate random signals to the exposure control signal on an electron-beam or laser-beam direct-write system. The only advantages of this technique over previous direct-written diffractive optics are that the complexity of the design procedure is simplified and the number of values placed in machine memory can be greatly reduced.

C. Alternative Implementations and Applications of Diffuser Arrays

We briefly mention two other potential applications of the diffuser array concept. Polymer-dispersed liquid crystal under applied voltage can be converted between isotropic and randomly oriented states. It may be possible to develop a real-time SLM in which this type of liquid-crystal layer is cascaded with pure-phase-retarding pixels. We present this device more as an illustration of the concept of diffuser arrays than as a serious candidate device. The currently prevailing view is that the development of any tandem SLM is too costly and risky. The second application is to use patterned diffuser arrays as gray-scale masks in projection printers. These masks could be used in place of true gray-scale and halftone masks that were recently used to demonstrate projection printing of three-dimensional diffractive optical structures in photoresist.[17–19] For either the halftone mask or the pseudorandom patterns, grayscale is achieved by diffracting light outside the aperture of the imaging lens. Speckle will not be present in the projected image if the source illumination is adequately incoherent. The gray-scale effect can be easily demonstrated by placing a piece of ground glass on the platen of an overhead projector. The pseudorandom masks for projection printing could be fabricated with either system proposed in Fig. 4. The remainder of this paper considers technical issues associated with the patterning systems in Fig. 4.

4. TECHNICAL CONSIDERATIONS FOR PATTERNING DIFFUSER PIXELS IN PHOTORESIST

A. Issue 1: Proximity Recording of Laser Speckle

Projecting laser speckle through a small aperture may unacceptably blur the exposure pattern. As an example, consider Fig. 5, which shows how a fully developed speckle pattern (457-nm argon-ion wavelength) diffracts at various distances past a 100-μm slit. At a distance of 100 μm past the slit, the edges of the pattern are still rather sharp, showing a transition from light to dark on the order of 10 μm. This indicates that pixels having a large fill factor can be made by proximity exposure for reasonable (10–100-μm) separations between mask and resist. It is even feasible to maintain separations of less than 10 μm, but at minute distances there would be little further reduction in the shadow region on account of the resist thickness, which may be exposed to a depth of several micrometers (described in Subsections 4.B–4.G).

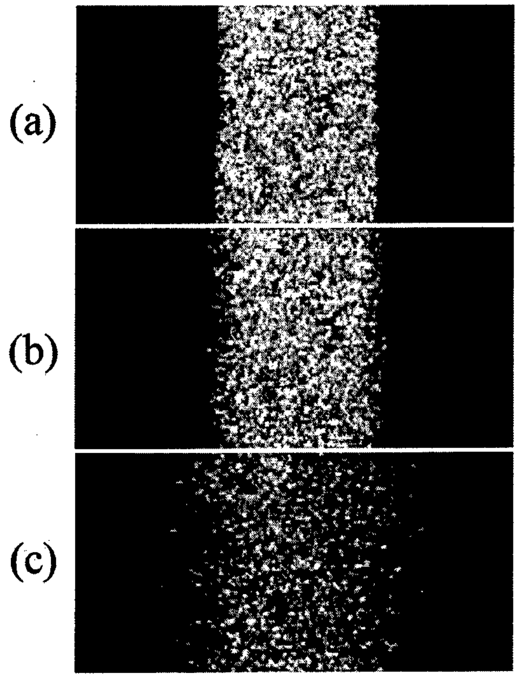


Fig. 5. Gray-scale intensity images of speckle patterns recorded at (a) 0 μm, (b) 100 μm, and (c) 500 μm past a 100-μm slit. The diameter of the speckle is approximately 2.5 μm. Patterns were imaged onto a 1/3-in. (0.85-cm) CCD camera by using a 40× microscope objective approximately 160 mm from the CCD. The images were then recorded with a video frame grabber.

As compared with recording interference fringes, speckle requires minimal vibration isolation. For a diffuser, a laser, and a CCD observation camera on a 2-in.-(5.08-cm-) thick optical bread board supported by a wood table, we observed that speckle patterns displayed on a video monitor exhibited no apparent displacement for speckle diameters larger than 2 μm. Vibration was noticeable for 0.6-μm speckle, but no blurring was observed for 1/30-s exposure frames recorded by using a frame grabber. Thus it seems that it is quite practical to illuminate resist with 2-μm speckle through an aperture in near contact (100 μm or less). For pixels of the order of the size of current SLM pixels (12.5–100 μm), the directivity gains can be 39 to 2500.

B. Issue 2: Complex Modulation for Recording Speckle in Linear Resist

The pdf of I_s, the intensity of fully developed speckle, is known to be exponentially distributed[1,2] and is written as

$$p(I_s) = \frac{1}{\langle I_s \rangle} \exp\left(\frac{-I_s}{\langle I_s \rangle}\right), \tag{7}$$

where $\langle I_s \rangle$ is the average intensity of the speckle pattern. Also, since speckle intensity is exponentially distributed, $\langle I_s \rangle$ can be interpreted as the standard deviation of the speckle intensity. For a photoresist that linearly maps exposure energy into resist thickness, ψ_s, the random phase depth produced, is proportional to exposure energy E_s and intensity I_s of the speckle pattern. Likewise, a mutually incoherent and spatially uniform illumination can be used to produce a bias phase shift ψ_b, so that the total random phase shift can be expressed as $\psi = \psi_b + \psi_s$, where ψ_b is proportional to E_b, the bias exposure. The effective complex modulation produced by this surface can be found by treating the actual phase depth ψ as an exponentially distributed random variable. Using the pdf for ψ of the form of Eq. (7) in Eq. (3) yields

$$\langle \boldsymbol{a} \rangle = \frac{\exp[j(\psi_b + \arctan\langle \psi_s \rangle)]}{\sqrt{1 + \langle \psi_s \rangle^2}}. \tag{8}$$

The amplitude decreases monotonically with increasing average phase depth of the resist, $\langle \psi_s \rangle$ (which is also proportional to average energy density of the speckle, $\langle E_s \rangle$). The phase shift that is due to speckle alone varies only from zero to $\pi/2$, but ψ_b can be chosen to produce any phase shift from zero to 2π.

C. Issue 3: Selecting Resist Thickness to Ensure Linearity

A linear resist will effectively saturate if developed through its entire thickness down to the substrate. This nonlinearity will change the complex modulation over that predicted by Eq. (8). Consider that the total resist thickness is proportional to the maximum phase shift $\psi_m = \psi_b + \psi_{ms}$, where ψ_{ms} is the maximum phase shift available for speckle recording at a given bias. The effective complex modulation for this case is found by evaluating Eq. (3) as

$$\langle \boldsymbol{a} \rangle = \exp(j\psi_b)\left[\int_0^{\psi_{ms}} p(\psi)\exp(j\psi)\mathrm{d}\psi + \exp(j\psi_{ms})\int_{\psi_{ms}}^{\infty} p(\psi)\mathrm{d}\psi\right], \tag{9}$$

where the density function is of the exponential form in Eq. (7). This evaluates to

$$\langle \boldsymbol{a} \rangle' = \frac{\exp[j(\psi_b + \arctan\langle \psi_s \rangle)]}{\sqrt{1 + \langle \psi_s \rangle^2}} \times \left\{1 + \langle \psi_s \rangle \exp[j(\psi_{ms} - \pi/2)]\exp\left(\frac{-\psi_{ms}}{\langle \psi_s \rangle}\right)\right\}, \tag{10}$$

where the prime is used to indicate that this result is perturbed from the result in Eq. (8). If the saturated value of phase ψ_{ms} is much greater than $\langle \psi_s \rangle$, the average phase produced by a purely linear recording of speckle, then Eq. (10) reduces to Eq. (8). Thus the second term in braces in Eq. (10) represents the errors that are due to finite resist thickness. A minimum thickness can be selected based on the minimum amplitude $a_{\min}$ of $a_c \in [a_{\min}, 1]$ that is practical to implement and the maximum allowable error ϵ between Eqs. (8) and (10). The worst-case absolute error is approximately

$$\epsilon(a_c) \approx \exp(-a_c\psi_{ms}), \tag{11}$$

where the approximation $a_c = |\langle \boldsymbol{a} \rangle| \approx 1/\langle \psi_s \rangle$ for average phase depth much greater than 1 rad has been used in Eq. (10). The minimum total resist thickness is then proportional to

$$\psi_t \equiv \psi_{mb} + \psi_{ms} = 2\pi - (\ln \epsilon_{\min})/a_{\min}, \tag{12}$$

where $\epsilon_{\min} = \epsilon(a_{\min})$ and $\psi_{mb} = 2\pi$ is the maximum bias shift required to achieve all possible phase shifts. Using ψ_{ms} as defined in Eq. (12) in relation (11) gives error as a function of a_c of

$$\epsilon(a_c) = (\epsilon_{\min})^{a_c/a_{\min}}. \tag{13}$$

As a specific example of using these equations to select resist thickness, consider the case for a minimum amplitude of $a_{\min} = 0.025$ and an absolute error of $\epsilon_{\min} = 0.0025$, or a 10% relative error. With the use of Eq. (12), the resist thickness is $\psi_t = 246$ rad or 39.1 optical wavelengths. For a reflective surface relief pattern and an optical wavelength of 0.633 μm, the resist can be as thin as 12.3 μm. Equation (13) shows that the relative error decreases rapidly for $a_c > 0.025$. For example, for $a_c = 0.03$ the error drops to 0.00075. Resist thickness is then only a significant concern for very small amplitudes, i.e., those values smaller than 0.025. The thickness is quite reasonable for standard photoresists.[20,21]

For comparison, the Kirk–Jones method using a sinusoidal carrier requires a thickness of at least

$$\psi_t = 2\pi + 2J_0{}^{-1}(a_{\min}), \tag{14}$$

which follows from Eqs. (5) and (6). For $a_{\min} = 0$ the total thickness for a reflective surface is 0.56 μm. While the thickness of the resist for the random method is much larger than that for the deterministic method, it should be recognized that the selection of thickness in relation (11) and Eq. (12) used a worst-case design. Furthermore, the maximum average speckle exposure energy is proportional to $\langle \psi_s \rangle \approx 1/a_{\min}$, which corresponds to an average depth of 2 μm. Thus the comparison in terms of energy use is more favorable. The basic conclusion for these numerical examples, is that the resist can be treated as infinitely thick for resists six times thicker than the average speckle depth.

The pseudorandom method can also produce an effective zero. If the magnitude of the second term in braces in Eq. (10) is unity and $\psi_{ms} = \psi_t - \psi_b$ and $\langle \psi_s \rangle$ are chosen to produce a phase shift of π, then Eq. (10) is zero. This is equivalent to having a relative error of 100% between Eqs. (8) and (10). For example, for the 39.1-wavelength-thick resist discussed above, an exposure depth of 9.6 wavelengths or 3.05 μm produces a zero according to Eq. (10) as compared with an amplitude of $a_c = 0.0165$ for an infinitely thick resist [according to Eq. (8)]. Unless the exposure system is precisely controlled and the resist thickness is precisely known, it would actually be quite difficult to implement a true zero accurately by this method. In most applications a very low minimum effective amplitude should be adequate.

D. Issue 4: Transformation of Speckle Statistics by Recording in Log Nonlinear Resist

For many resists, thickness is proportional to the logarithm of exposure over a wide dynamic range. For such resists the exposure curve (depth into the resist, t, versus exposure energy E) takes the form

$$t(E) = m\ \ln(E/E_b), \tag{15}$$

where E_b is a reference recording level corresponding to a reference thickness of $t = 0$ and m is the logarithmic slope of the resist. The exposure curve of a 9.5-μm-thick film of resist (AZ 4903 positive) presented in Ref. 16 is well fit over a 7-μm range for a slope of $m = 2.70\ \mu$m and a reference energy of $E_b = 75$ mJ/cm^2. For 5-μm films of Shipley S1650 resist, we have experimentally determined that the slope is $m = 0.823\ \mu$m over a 2.6-μm range starting from a reference energy of $E_b = 40$ mJ/cm^2.

Using the logarithmic range of a resist leads to an effective amplitude that depends on the ratio of speckle exposure to bias exposure rather than absolute intensity. This may prove to be an advantageous feature, since it is often easier to control ratios (using a half-wave plate and a polarized beam splitter) than it is to control the absolute energy individually in two independent exposures.

The effective amplitude can be found by using the following analysis. The logarithmic recording medium produces the total phase shift

$$\psi_t = \psi_b + \psi_s = \alpha\ \ln(E_s + E_b), \tag{16}$$

where α is the logarithmic slope in radians (i.e., $\alpha = 4\pi m/\lambda$ for a reflective surface) and $\psi_b = \alpha\ \ln(E_b)$. This definition allows the phase shift that is due to speckle to be written as

$$\psi_s = \alpha\ \ln(1 + E_s/E_b). \tag{17}$$

Using the definitions in Eqs. (15) and (16), the exponential density of the form of Eq. (7), and the change of variables $x = E_s/\langle E_s\rangle$ in Eq. (3) leads to

$$\begin{aligned}\langle \boldsymbol{a}\rangle &= \exp(j\psi_b)\int_0^\infty \exp(-x)\exp[j\alpha\ \ln(1+\gamma x)]\mathrm{d}x\\ &= \exp[j(\psi_b + \alpha\ \ln\ \gamma)]\exp(1/\gamma)\Gamma(1+j\alpha,\ 1/\gamma),\end{aligned} \tag{18}$$

where $\gamma = \langle E_s\rangle/E_b$ and $\Gamma(a,\ b)$ is the incomplete gamma function.[22] Figure 6 shows the effective amplitude produced by exposing the S1650 and AZ4903 resists (described above) with speckle patterns and then reflecting 633-nm light from the resulting surfaces. This corresponds to using $\alpha = 16.34$ and 53.6 in the evaluation of Eq. (18). For these values of α, the effective phase (excluding bias ψ_b) varies by slightly more than $\pi/2$ for all values of γ. This amount of phase modulation is comparable with the maximum phase shift for linear resists [see Eq. (8) and Fig. 6].

The minimum resist thickness that effectively behaves as infinitely thick [thus permitting the use of Eq. (18)] can be determined by an analysis similar to that in Subsection 4.C. The maximum phase shift for which the resist is exposed down to the substrate is once again written as $\psi_m = \psi_b + \psi_{ms}$. However, the maximum phase shift that is due to speckle is now explicitly written as $\psi_{ms} = \alpha\ln(1 + E_{ms}/E_b)$, where E_{ms} is the amount of energy above bias at which the resist is completely exposed. With these definitions the perturbed version of Eq. (18) is written as

$$\begin{aligned}\langle \boldsymbol{a}\rangle' &= \langle \boldsymbol{a}\rangle + \epsilon\\ &= \langle \boldsymbol{a}\rangle + \exp(j\psi_b)\left\{\exp\left(\frac{-\gamma_{ms}}{\gamma} + j\psi_{ms}\right)\right.\\ &\quad \left. - \int_{\gamma_{ms}/\gamma}^\infty \exp[-x + j\alpha\ \ln(1+\gamma x)]\mathrm{d}x\right\},\end{aligned} \tag{19}$$

where the definition $\gamma_{ms} = E_{ms}/E_b$ has been used and ϵ is the absolute error resulting from the perturbation.

Continuing with the numerical example begun in Subsection 4.C, a value of γ is found, by using Eq. (18), for which $a_c = 0.025$. For the resist with the smaller logarithmic slope ($\alpha = 16.34$), a value of $\gamma = 2.45$ is needed to produce this amplitude. For the resist with the larger slope, a value of $\gamma = 0.745$ is needed. For an absolute error $\epsilon < 0.0025$, then, the ratio $\gamma_{ms}/\gamma = E_{ms}/\langle E_s\rangle$ needs to be approximately 6 or greater [as found by numerical evaluation of Eq. (19)]. This is essentially identical to the result for linear photoresist. However, on account of the nature of the logarithmic resists, the resist thickness can be much less than that for linear resists. The minimum thicknesses are 2.26 μm for the low-α resist and 4.59 μm for the high-α resist, as compared with 12.1 μm for the linear resist. The required thickness can be appreciated by comparing it with the pdf for the recorded depths (which are proportional to the random phases ψ_s). This is shown in Fig. 7. The density function for logarithmically recorded speckle has been derived by a standard technique for transformations of random variables.[4] This function is written as

$$p(\psi_s) = \frac{1}{\alpha\gamma}\exp\left\{\frac{\psi_s}{\alpha} + \frac{1}{\gamma}\left[1 - \exp\left(\frac{\psi_s}{\gamma}\right)\right]\right\}. \tag{20}$$

Note that for each curve in Fig. 7, $p(0) = 0.025 = a_c$. Also note that for the logarithmic resists, $p(0) = 1/(\alpha\gamma)$. The relationship between the pdf and the effective ampli-

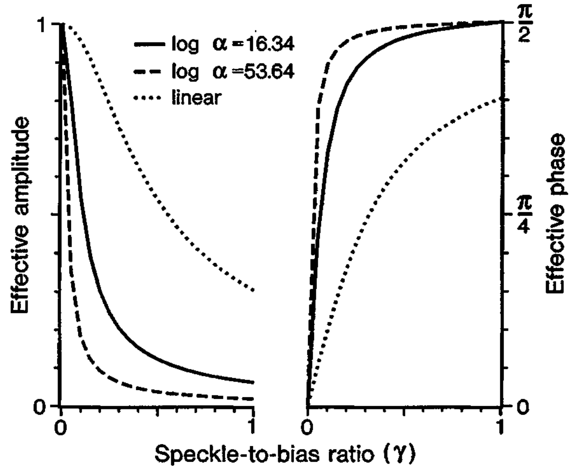


Fig. 6. Effective amplitude a_p and phase ϕ_p for log and linear resists. For linear resist, which depends on absolute intensity, the x axis is defined to be $\gamma = \langle\psi_s\rangle/\pi$.

tude is approximately valid for $\alpha \geqslant 4$. For α near 5 the pdf curve is even more sharply peaked and narrower than that for the $\alpha = 16.34$ curve, and the maximum resist thickness is approximately 1 μm. For $\alpha > 53.6$ the pdf more closely approaches the exponential distribution for a linear resist. For an appropriately chosen value of α, a logarithmic transformation of speckle permits the use of much thinner films than those for linear resists.

E. Issue 5: Special Case: Low-Sensitivity Log Resist

For resists having sensitivities below 4, the effective amplitude cannot be continuously controlled between one and zero. This can be seen by evaluating Eq. (18). For large values of γ, the effective amplitude is well approximated as

$$\langle \boldsymbol{a} \rangle \approx \exp[j(\psi_b + \alpha \ln \gamma)]\Gamma(1 + j\alpha), \qquad \text{(18a)}$$

where $\Gamma(\cdot) \equiv \Gamma(\cdot, 0)$ is the gamma function. The magnitude of relation (18a) decreases monotonically with increasing α. For example, for $a_p = 0.01, 0.25, 0.5$, and 0.75, $\alpha = 4, 1.62, 1.04$, and 0.625, respectively. The effective amplitude as a function of γ [as calculated by using Eq. (18)] can oscillate around the limiting value of effective amplitude, but this is usually a negligible amount. The only significant undershoot is evident for α close to $\alpha = 2.72$. In this instance the effective amplitude as a function of γ dips to zero (at $\gamma \simeq 5$) before settling to an effective amplitude of 0.058. The most important point is that a high-contrast material ($\alpha > 4$) is required in order to produce fully complex modulation.

F. Issue 6: Time-Averaged Recording in Linear Resists

The patterning system in Fig. 4(a) uses time averaging of speckle patterns (achieved by varying the velocity of the spinning diffuser) to control both effective amplitude and step height. The effect of time averaging of speckle patterns is reasonably modeled as the addition of M equal-intensity exposures of uncorrelated speckle patterns in sequence (Ref. 1, Chap. 4). With exposure time and exposure energy held constant, the parameter M is proportional to the velocity of the diffuser. Alternatively, with velocity held constant, M is proportional to exposure time. Except for values of M close to unity, the resulting curves for effective amplitude can be accurately interpolated for continuous values of M.[1] Here we model time-averaged complex recording in linear resist. Then, in Subsection 4.G we consider time averaging in nonlinear resist.

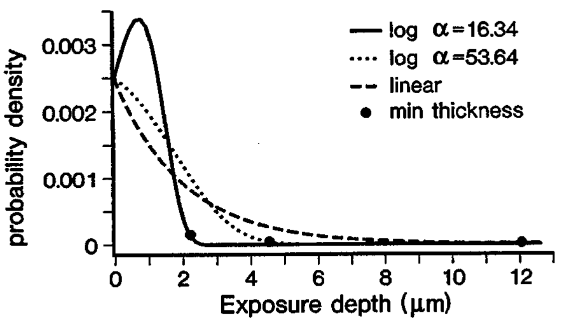


Fig. 7. Probability density functions for depths of speckle recorded into log and linear resists. Each distribution produces effective amplitude $a_p = 0.025$.

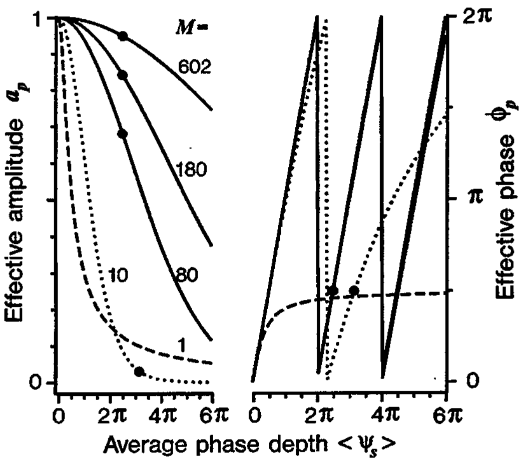


Fig. 8. Effective complex amplitude for time-averaged recording of speckle in linear resist. Average recorded depth is proportional to average exposure energy $\langle E_s \rangle$. The amplitude and phase curves use the same style for a given value of M. For M of 80, 180, and 602, the effective phase curves are nearly identical and, for this reason, are plotted with a single style. The dots (●) indicate where effective phase is 2.5π.

The pdf for each exposure is Eq. (7), and the pdf for the total exposure is the result of convolving the M identical pdf's.[4] The pdf for the phase ψ_s that is due to this total exposure is the gamma density[1]

$$p(\psi_s) = \frac{{\psi_s}^{M-1}}{\Gamma(M)} \left(\frac{M}{\langle \psi_s \rangle} \right)^M \exp\left(-\frac{M\psi_s}{\langle \psi_s \rangle} \right), \qquad \text{(21)}$$

where $\langle \psi_s \rangle / M$ is proportional to the exposure energy of an individual speckle pattern. The effective complex amplitude is known to be the characteristic function of the pdf evaluated at frequency equal to unity,[2] and thus the complex amplitude is of the form of the Mth power of Eq. (8):

$$\langle \boldsymbol{a} \rangle = \left[1 + \left(\frac{\langle \psi_s \rangle}{M} \right)^2 \right]^{-M/2} \exp\left[j\left(M \arctan \frac{\langle \psi_s \rangle}{M} \right) \right]. \qquad \text{(22)}$$

The effective amplitudes and phases of Eq. (22) are plotted in Fig. 8 against average exposure and for various values of M. The dots on the curves indicate specific points for which the effective phase shift is 2.5π. For the dot markers the amplitude varies between 0.031 and 0.95 for M between 10 and 602. [For $M = 1$ the results are the same as those with Eq. (8).] Near-unity amplitudes can be produced, but not for all values of phase. It may not be practical to increase M further, as this increases recording time. One way to address this wide variation in M is to control multiple parameters such as intensity, diffuser angular velocity, and radial position of the laser beam on the diffuser. This would allow a modest range of control (less than 10:1) on each of the three parameters. It may also be desirable to add a separate phase bias ψ_b for amplitudes that are close to unity in order to reduce recording time.

The principal advantage of time-averaged recording is that the maximum recording depth is substantially less than that for nonaveraged recording. This is shown in Fig. 9 for the density functions corresponding to $M = 2$, 10, and 29 and effective amplitude $a_p = 0.025$. The effective values of phase shift are, respectively, 0.9π, 2.6π, and 4.6π. These curves can be compared with Fig. 7. They are substantially narrower than the exponential density. The curves for $M = 1$ to 10 and $M = 10$ to 29 both produce a 2π range; however, the second set of curves (compare $M = 2$ with $M = 29$) are even narrower. Also, the exposure energy used for time-averaged recording will be smaller by a factor of 2 to 4. This can be seen by inverting the amplitude in Eq. (22) for average exposure energy:

$$\langle E_s \rangle \propto \langle \psi_s \rangle = M\sqrt{a_p^{-2/M} - 1}. \tag{23}$$

For $a_p = 0.025$ and $M = 1$, which corresponds to the exponential distribution, the average intensity is proportional to 12.7π. For $M = 2$ the exposure drops to 4.0π. For $M = 5$ the energy is minimum at 2.9π, and it increases gradually to 5.0π at $M = 29$. Therefore both exposure energy and film thickness can be much less if temporal averaging is used.

For monochromatic diffractive optic design, phase modulations ϕ_p that differ by integer multiples of 2π are often treated as equivalent. This 2π phase ambiguity can be used to produce the same effective value of the complex modulation $\boldsymbol{a}_c = (a_c, \psi_c)$ for different exposure conditions. A particular choice of exposure conditions may be preferable from various considerations of energy efficiency, accuracy, and recording time. We illustrate this by expressing the amplitude a_p in terms of the phase ϕ_p in Eq. (22). The term $\langle \psi_s \rangle / M$ that is in common between the expressions for amplitude and phase is substituted out to give

$$a_p = |\cos(\phi_p / M)|^M. \tag{24}$$

Figure 10 plots both the desired amplitude [Eq. (24)] and $\langle \Psi_s \rangle$ (which is proportional to average exposure energy) against M (which is proportional to the amount of time averaging). The three curve styles are used to distinguish the results for three effective values of phase ϕ_p that differ from each other by integer factors of 2π (specifically, $\pi/2$, $5\pi/2$, and $9\pi/2$). The solutions represented by dots in Fig. 8 (for $M = 10$, 80, and 180) are replotted in Fig. 10 (again shown as dots). These values were calculated for $\phi_p = 5\pi/2$ and thus are located on the solid curves. For each of these solutions, there is an alternative recording condition that produces the same complex amplitude (indicated in Fig. 10 by diamonds).

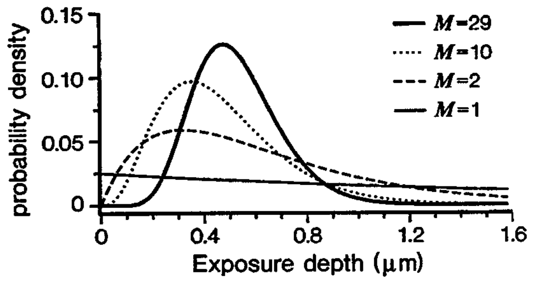


Fig. 9. Probability density functions (pdf's) for time-averaged recording in linear resist. Each pdf produces identical effective amplitude $a_P = 0.025$.

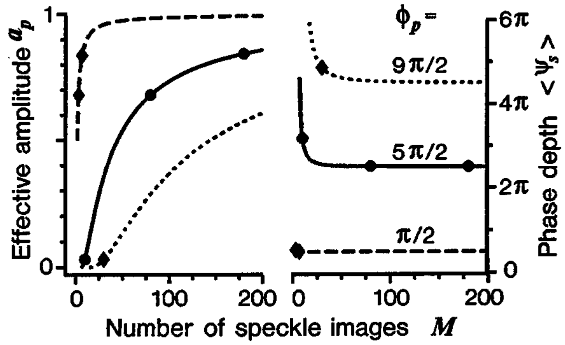


Fig. 10. Effective amplitude for time-averaged recording in linear resist for a constant value of effective phase. The dots (●) indicate identical points from Fig. 8. The diamonds (♦) indicate points identical in amplitude but differing in phase by an integer multiple of 2π.

We can see that one recording condition may be preferred to another. For instance, for the smallest effective amplitude ($a_p = 0.031$) the solution on the $9\pi/2$ curve uses more energy than that for the $5\pi/2$ curve, but the amplitude is less sensitive to exposure time. For the two larger-amplitude solutions, the alternative choices on the $\pi/2$ curve use less energy but are much more sensitive to exposure time. Figure 8 also shows that the sensitivity of the amplitude with respect to exposure energy generally decreases with increasing exposure energy.

For large amounts of time averaging, Eq. (22), the effective amplitude, can be simplified. The gamma density function in Eq. (21) can be approximated as a Gaussian of the form

$$p(\psi_s) \approx \frac{1}{\sqrt{2\pi M}\langle \psi_s \rangle} \exp\left[\frac{1}{2M}\left(\frac{\psi_s - \langle \psi_s \rangle}{\langle \psi_s \rangle}\right)^2\right]. \tag{21a}$$

for M a large number, through the use of the central limit theorem (see Ref. 4, pp. 214–221 and 240). Substituting this result in Eq. (3) approximates the effective amplitude of Eq. (22) as

$$\langle \boldsymbol{a} \rangle \approx \exp(j\langle \psi_s \rangle)\exp\left(\frac{-\langle \psi_s \rangle^2}{2M}\right). \tag{22a}$$

This result is quite good for $M > 10$. This result accurately describes the effective amplitude and phase for $M = 80$, 180, and 602. In particular, note that the effective phase ϕ_p is independent of M. This can be seen in Fig. 8, where the effective phase curves (and also the three dots at $\langle \psi_s \rangle = 2.5\pi$) are all nearly identical.

G. Issue 7: Time-averaged Recording in Log Resist

The analysis of effective amplitude is identical to that used in deriving Eq. (18), except that the gamma density is used in place of the exponential density. This gives

$$\langle \boldsymbol{a} \rangle = \exp(j\psi_b) \int_0^\infty \frac{x^{M-1}}{\Gamma(M)} \exp(-x) \times \exp\left[j\alpha \ln\left(1 + \frac{\gamma x}{M}\right)\right] \mathrm{d}x. \tag{25}$$

The amplitude again depends on the ratio of speckle intensity to bias intensity. For $M = 1$ Eq. (25) is identically Eq. (18). For any value of M, the amplitude decreases monotonically with increasing γ. For M a large number, the gamma density in Eq. (25) can be replaced by its approximate form [relation (21a)]. After an appropriate change of variables, Eq. (25) is approximated as

$$\langle \boldsymbol{a} \rangle \approx \frac{\exp(j\psi_b)}{\sqrt{2\pi}} \int_{-\infty}^{\infty} \exp(-x^2/2) \times \exp[j\alpha \ln(1 + \gamma + \gamma x/\sqrt{M})]\mathrm{d}x. \qquad (25a)$$

Factoring out the term $1 + \gamma$ in the log function and using the approximation $\ln(1 + z) \approx z$ for values of $z < 1$, we can further simplify Eq. (25) to

$$\langle \boldsymbol{a} \rangle \approx \exp[j\psi_b + j\alpha \ln(1 + \gamma)]\exp\left[\frac{-1}{2M}\left(\frac{\alpha\gamma}{1 + \gamma}\right)^2\right]. \qquad (25b)$$

The range of validity of the expansion depends on the extent of the Gaussian in relation (25a). The Gaussian is essentially zero for $x > 3$. This leads to $M > 9[\gamma/(1 + \gamma)]^2$, which is always true for $M > 9$.

Relation (25b) shows that the effective amplitude a_p monotonically decreases with increasing speckle-to-bias ratio γ. For low-sensitivity resist (see Subsection 4.E), the curves saturate without reaching zero. Increasing M raises only the saturation value and does not increase depth of amplitude modulation over recording without time averaging. The amplitude control provided by time averaging in log photoresist is similar to that for time-averaged recording in linear resists, as can be seen by comparing relations (22a) and (25b). The main difference between the two results is that relation (22a) always approaches zero given a large enough exposure, while relation (25b) instead settles to a constant amplitude determined by α^2/M.

H. Issue 8: Spatial Resolution of Linear and Log Resists

Photoresists generally have much higher spatial resolution than the diffraction limit. However, if speckle is reimaged through a projection system, it would be possible to use an adjustable iris in place of the spinning diffuser. The blurred speckle pattern can then be considered as spatially integrated. The problem is analyzed in Chap. 2 of Ref. 1, and it is not surprising that the results are identical to the analysis of time-integrated speckle given above. As above, the gamma function is a good approximation of the pdf of the spatially averaged speckle intensities. The parameter M is now interpreted as the effective number of speckles averaged together in a rectangular window. Thus the results presented above in Subsections 4.F and 4.G can be used without modification to analyze the effect of resolution loss in linear and logarithmic resists.

5. EXPERIMENTAL DEMONSTRATION OF SPECKLE RECORDING

The theory presented in Section 4 primarily describes the complex amplitudes that could be produced by recording laser speckle in photoresist. In order to anticipate better the potential problems in developing the proposed exposure system, we have also performed some preliminary experiments in which we use a phase-only liquid-crystal light valve to represent photoresist. Unlike the demonstration reported in Section 2, in which the SLM represented an array of pixels, in this section the SLM represents a single pixel.

One purpose of the demonstration is to show experimentally the control of effective amplitude by varying the exposure patterns. In one set of measurements, the exposure energies of speckle, $\langle E_s \rangle$, and of bias, E_b, are varied. This corresponds to exposure using the apparatus in Fig. 4(b). In a second set of measurements, the speckle energy and the speckle diameter are varied. The SLM has limited spatial resolution, so the SLM introduces spatial averaging. As pointed out in Subsection 4.H, spatial averaging gives results that are mathematically equivalent to time averaging. Thus this second set of measurements is representative of results made possible by using the patterning system in Fig. 4(a).

The second purpose of the demonstration is to relate the experimental measurements to our theory of speckle recording. However, the optical characteristics of SLM's that we have studied are much more complicated than the properties assumed for resists. The SLM used for this demonstration is a gallium arsenide photodetector, birefringent liquid-crystal light valve from the Lebedev Physical Institute, Moscow. It was chosen because it produces the largest phase shift (up to 4 π) of the SLM's available to us. Measurements in a Michelson interferometer of the read side of the light valve indicate that there is a roughly logarithmic dependence of the phase modulation depth on the exposure intensity. However, the exact phase shifts measured can vary dramatically based on the spatial-frequency content of the illumination and the exposure intensity. In particular, the spatial resolution of the device (4 to 40 line pairs per millimeter) is known to depend on the exposure intensity. Rather than attempting to measure and then model the SLM completely, we have attempted empirically to fit theoretical curves for a logarithmic film of resist to the measured response of the SLM. This is to say that it has been possible to adjust parameters (specifically, α, E_b, ψ_{ms}, and M) in the theoretical equations so that the trends in the experiment match the theory. This exercise is certainly valuable for better appreciating the theory and for anticipating the practical limitations of actual photoresists. We also believe that the process of comparison of the experiment with an approximate theory provides insight into the optical characteristics of the SLM.

A. Measurement Procedure

The effective amplitude is measured by using the following procedure. The write side of the light valve is illuminated by two mutually incoherent (850-nm) laser diode sources. One beam is expanded and illuminates the light valve with a spatially uniform bias. The other beam is focused into a small spot on the surface of a ground-glass diffuser to produce a speckle pattern illumination on the light valve. The speckle diameter is varied by translating a diffuser along the path of the beam so as to change the beam diameter intercepting the diffuser. The light valve is electrically driven with a 2-kHz, 10-V rms sinusoidal potential from a signal generator. The read side of the light valve is illuminated with a 633-nm-wavelength HeNe laser beam. The beam is spatially filtered and expanded by using a collimator. The collimator lens is positioned to converge the beam slightly. At the face of the SLM, the beam is 14.5 mm in diameter. The reflected beam is observed by using a CCD camera positioned at the focus of the collimator lens. A digital oscilloscope connected to the video output of the camera is used to measure the intensity of the specular diffraction peak for a range of speckle exposure (0–365 μW/cm^2) and different settings of bias exposure (0, 9.8, 15.3, and 29.3 μW/cm^2) or speckle diameter (0.07, 0.15, 0.25, 0.4, 1, and 3 mm). The measured intensities are normalized so that the SLM has nominally unity transmittance for zero intensity exposure. The measured effective amplitude a_p is taken to be the square root of the normalized intensity. These results are plotted in Fig. 11. The results for speckle recording at different levels of bias [Fig. 11(a)] and for various speckle diameters [Fig. 11(b)] are discussed in sequence.

B. Complex Recording by Combined Speckle and Bias Exposure

As discussed in Subsection 4.D, for an ideal logarithmic resist of infinite thickness, Eq. (18) shows that the ratio of speckle energy to bias energy, $\gamma = \langle E_s \rangle / E_b$, determines the effective amplitude and that the bias energy can be used to offset the effective phase by $\psi_b = \alpha \ln (E_b)$. This result becomes complicated for resists for which the exposed depth approaches the film thickness, as modeled by Eq. (19). For a thin film the total phase modulation range is $\psi_m = \psi_b + \psi_{ms}$, where ψ_{ms} is the phase modulation range available in the resist for speckle exposure. If ψ_{ms} is too small, then the effective amplitude cannot be varied from unity to zero. Thus, for a fixed thickness film, the amplitude range should decrease as the bias is increased. We will refer to this effect as *saturation* of the effective amplitude.

This saturation is observed for the measured curves in Fig. 11(a). For each curve the amplitude decreases with increasing speckle level to a point and then begins increasing. As the bias level (listed in Table 2) is increased, the minimum amplitude increases correspondingly. The bias levels have been selected so as to produce phase shifts ψ_b (these values, which were measured in the Michelson interferometer, are listed in Table 2) covering a 2π range. Thus this SLM, while it can control amplitude over a 10:1 range, cannot simultaneously produce all values of phase. Basically, the SLM needs more (than its current 4π) phase modulation range to achieve arbitrary phase and a 10:1 amplitude control.

For an ideal logarithmic resist, the effective amplitude is governed by Eq. (18) for low combined levels of speckle and bias exposure. Thus plots of effective amplitudes versus speckle-to-bias ratio γ will appear identical for low-level exposures. For the measured curves in Fig. 11(a), the initial slopes differ if the measured values of E_b are used. In order to compare the measurements for the SLM with the theory for an ideal resist, values of E_b (listed in the "E_b theory" column in Table 2) are used to calculate γ for the measured curves in Fig. 11(a). With these choices of E_b, the initial slopes of curves 1 and 4 are brought into coincidence with curves 2 and 3 (which are plotted by using the measured values of E_b). With these adjustments it is possible to compare the experimental results with the theory for logarithmic resist.

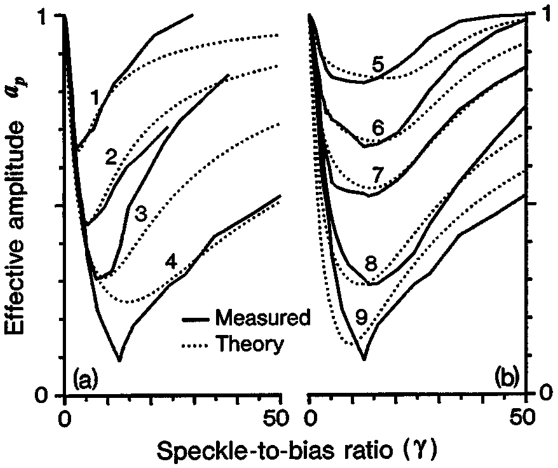


Fig. 11. Experimental demonstration of speckle recording using a phase-only liquid-crystal light valve to represent a photoresist. The plots show how the effective amplitude curves change for (a) different levels of uniform bias E_b and (b) different speckle diameters. Specific values used in the experiment and the theory are given in Tables 2 and 3.

Table 2. Parameters Used for the Measured and Theoretical Curves in Fig. 11(a)

	E_b (μW/cm^2)		ψ_b (rad)	
Curve ID	Measured[a]	Theory[b]	Measured[a]	Theory[b]
1	29.3	12.3	2.0π	0.91π
2	15.3	15.3	1.5π	1.26π
3	9.8	9.8	1.0π	1.67π
4	0.0	5.1	0.0π	2.06π

[a] Measured speckle diameter is 1 mm.
[b] Theoretical value of α is 1.65 rad.

The theoretical curves in Fig. 11(a) are calculated by using Eq. (19). The value of logarithmic slope $\alpha = 1.65$ has been selected so that the initial slopes of theoretical and measured curves match. Then values of ψ_{ms} (the phase modulation range available for speckle recording, which is listed in Table 2) are selected to introduce the saturation effect near the minima of the measured curves. The minima of theoretical curve 4 cannot be brought much lower for any value of ψ_{ms} unless the sensitivity α is also increased, as described in Subsection 4.E. [A fit to curve 4 using a larger value of α will be described in discussing Fig. 11(b)]. The theory, while not in close agreement for large values of γ, does show the same upward trend with increasing saturation.

The values of ψ_b and ψ_{ms} in Table 2 also give an idea of the discrepancy between the ideal resist and the measured curves. If the SLM were to fit the model of the log resist closely, then we would expect that the total phase range of the resist, $\psi_m = \psi_b + \psi_{ms}$ (the sum of the last two columns of Table 2), would be a constant for each level of bias, rather than between 2π and 3π. As described in Subsections 4.D and 4.E, it is desirable to choose resists that are thick enough to avoid saturation and sensitivities large enough to allow an adequately large range of effective amplitude. The results in Fig. 11(a) illustrate the consequences of not meeting these conditions. We continue these comparisons for the recording of spatially averaged speckle.

C. Complex Recording by Spatial Averaging of Speckle

The measured results in Fig. 11(b) show how the effective amplitude changes as a function of γ for speckle of various diameters. No bias exposure is applied, but for purposes of comparing these results with Fig. 11(a), the same theoretical value of bias $E_b = 5.1\ \mu W/\text{cm}^2$ is used to plot the results. Note, in particular, that curves 4 and 9 are the same measurements. Curves 5–8 show a decreasing range of amplitude modulation as the speckle diameter (listed in Table 3) decreases. We presume that this is due to spatial averaging caused by the limited resolution of the SLM. Further evidence of this is that a curve for 3-mm-diameter speckle (not shown) is nearly identical to curve 9 for 1-mm speckle over most of the range of γ. The only apparent discrepancy is near the minimum of each curve, where the 3-mm case dips only to 0.16 instead of 0.09. We believe that this difference is due mainly to the increased level of background noise for the 3-mm case, which is anticipated as a direct result of its lower directivity (18:1 for the 3-mm case as opposed to 165:1 for the 1-mm case). The remainder of this subsection compares these results with those predicted for recording of spatially averaged speckle in logarithmic resist.

The model developed in Subsections 4.G and 4.H for recording spatially averaged speckle in logarithmic resist assumes that M speckles within a rectangular window are averaged together. This leads to the relationship that M is inversely proportional to the square of the speckle diameter. For the SLM the averaging mechanism is more complicated. We know that resolution is intensity dependent and that the spatial averaging mechanism is likely to differ from that of rectangular averaging. Nonetheless, for purposes of comparing the measurements with the theory, we will compare M with measured speckle diameter through the inverse square relationship.

The equation used to calculate the theoretical curves in Fig. 11(b) is not explicitly presented. It combines the results for thin logarithmic resists [from Eq. (19)] with the results for time-averaged resists [Eq. (25)], and it can be derived directly by using the gamma density function for $p(\psi)$ in Eq. (9). This equation is fitted to the measured curves by first fitting curve 7 as closely as possible by adjusting parameters ψ_{ms}, α, and M and then by adjusting only M to fit the four other curves. The values used for curve 7 are $\psi_{ms} = 2.26\pi$ (corresponding to a single value of film thickness), resist sensitivity $\alpha = 2.0$, and $M = 3$. The values of M for the four other curves are listed in Table 3. The values of M are related to the theoretical values of speckle diameter (also listed in Table 3) by selecting the measured and theoretical diameters to be the same for $M = 3$ and scaling the other values according to the inverse square relationship.

With respect to the experimental curves, curves 8 and 9 appear to be overly compressed and curves 5 and 6 appear to be overly expanded along the γ coordinate. It appears that the effective amplitude of the SLM is saturating more rapidly for increasing values of γ and M than the theory predicts. The values of measured and theoretical speckle diameter in Table 3 indicate that the SLM sensitivity decreases more rapidly with speckle diameter than does the model for the resist.

We also compare theoretical curves 4 and 9 in Figs. 11(a) and 11(b), respectively. Note that curve 9, which has a higher sensitivity ($\alpha = 2$) than that of curve 4 ($\alpha = 1.65$), also produces a lower minimum effective amplitude. If we were trying to fit only measured curve 9, then we would also need to select a somewhat lower value of bias E_b in order to match more closely the initial slope of measured curve 9. These results give additional insight into how the theory depends on the model parameters.

Table 3. Parameters Used for the Measured and Theoretical Curves[a] in Fig. 11(b)

Curve ID	Speckle Diameter (mm) Measured	Speckle Diameter (mm) Theory	M (Theory)
5	0.07	0.14	10.0
6	0.15	0.20	4.5
7	0.25	0.25	3.0
8	0.4	0.35	1.5
9	1.0	0.43	1.0

[a] Measured bias is $E_b = 0\ \mu\text{W/cm}^2$; γ for the measured curves is calculated by using the theoretical value of bias $E_b = 5.1\ \mu\text{W/cm}^2$, and the theoretical curves by using α=2.0 rad and $\psi_{ms} = 2.26\pi$.

D. Summary of These Results

While the optical properties of the SLM and the idealized resist are quite different, similar trends are apparent. As discussed in Subsection 4.E, low values of sensitivity α limit the minimum achievable value of effective amplitude for a logarithmic resist; and, as discussed in Subsection 4.D, a finite phase modulation range ψ_{ms} causes the effective amplitude to increase for large-intensity speckle exposures. In fact, the phase modulation range is so small that any level of bias exposure at all reduces the total range of effective amplitude modulation. These characteristics seem also to describe qualitatively the behavior of the SLM, which we know has low (also signal-dependent) sensitivity and phase modulation range. For practical recording of arbitrary complex values, we clearly need greater phase range and sensitivity, especially since applying any bias (which is intended to realize the correct phase) further reduces the range of the effective amplitude. Likewise, speckle averaging reduces the depth of modulation, which limits our ability to achieve all complex values. These limitations reflect the shortcomings of using SLM's as demonstration vehicles, rather than of the concept of speckle recording itself. As described in Subsection 4.D, there are many resists that are adequately sensitive and that can be spun on in adequately thick layers.

6. SUMMARY AND CONCLUSIONS

In this paper we have presented the concept of the patterned diffuser array, in which desired complex-valued samples of a modulation function are realized as arrays of custom diffusers and where each diffuser pixel has an individually specified roughness and step height corresponding to the amplitude and the phase desired. The main application of this device is the realization of complex-valued spatial filters (e.g., composite pattern recognition filters, spot array generators, and structured light illuminators) with phase-only structures. A second potential application of patterned diffuser arrays is as gray-level photomasks for projection printing.

We have proposed a photoresist exposure system for the custom fabrication of diffuser arrays by exposing individual pixels to appropriate combinations of spatially uniform and nonuniform illumination. We have focused on and evaluated the feasibility by using speckle patterns (that occur naturally when a laser beam is passed through a diffuser) as the illumination source of the pattern generator. This exposure system appears to place no critical requirements on optical components, vibration isolation, or air cleanliness. For this reason we believe that the components required to construct a turnkey system would cost well under $100,000. The most costly component appears to be the translation stages, which should be as fast as possible to reduce fabrication time. If multiple copies of a diffuser array are required, then greater speeds are possible by using various replication methods.[8]

Patterned diffuser arrays provide a direct way to implement complex-valued modulation without resorting to numerically intensive design procedures. This approach could be used to shorten significantly the time required to design and, in many cases, to fabricate, a wide variety of diffractive optics functions.

ACKNOWLEDGMENTS

We thank K. M. Walsh of the University of Louisville for advice on photoresist properties and processing methodology. We also thank Hughes-JVC, Carlsbad, Calif., for the loan of the CRT and the optics for addressing the Hughes light valve and A. V. Parfenov of the Lebedev Physical Institute for the loan of the gallium arsenide light valve. This research was sponsored by the Advanced Research Projects Agency through Rome Laboratory contract F19628-92-K0021, U.S. Army Research Office contract DAAH04-93-G-0467, and National Aeronautics and Space Administration cooperative agreement NCCW-60 through Western Kentucky University.

*Permanent address, Department of Precision Instrument Engineering, Tianjin University, Tianjin, China 300072.

Address all correspondence to Robert W. Cohn at the address on the title page; tel: 502-852-7077; fax: 502-852-1577; e-mail: rwcohn01@ulkyvm.louisville.edu.

REFERENCES

1. J. C. Dainty, ed., *Laser Speckle and Related Phenomena,* 2nd ed. (Springer-Verlag, Berlin, 1984).
2. J. W. Goodman, *Statistical Optics* (Wiley, New York, 1985).
3. R. W. Cohn and M. Liang, "Approximating fully complex spatial modulation with pseudorandom phase-only modulation," Appl. Opt. **33**, 4406–4415 (1994).
4. A. Papoulis, *Probability, Random Variables, and Stochastic Processes*, 3rd ed. (McGraw-Hill, New York, 1991).
5. J. A. Davis and D. M. Cottrell, "Random mask encoding of multiplexed phase-only and binary phase-only filters," Opt. Lett. **19**, 496–498 (1994).
6. L. G. Hassebrook, M. E. Lhamon, R. C. Daley, R. W. Cohn, and M. Liang, "Random phase encoding of composite fully-complex filters," Opt. Lett. **21**, 272–274 (1996).
7. R. W. Cohn and M. Liang, "Pseudorandom phase-only encoding of real-time spatial light modulators," Appl. Opt. **35**, 2488–2498 (1996).
8. M. T. Gale, M. Rossi, J. Pedersen, and H. Schutz, "Fabrication of continuous relief micro-optical elements by direct laser writing in photoresists," Opt. Eng. **33**, 3556–3566 (1994).
9. W.-H. Lee, "Computer-generated holograms: techniques and applications," in *Progress in Optics*, E. Wolf, ed. (North-Holland, Amsterdam, 1978), Vol. 16, Chap. 3, pp. 119–232.
10. W. J. Dallas, "Computer-generated holograms," in *The Computer in Optical Research*, B. R. Frieden, ed. (Springer, Berlin, 1980), Chap. 6, pp. 291–366.
11. R. W. Gerchberg and W. O. Saxton, "Practical algorithm for the determination of phase from image and diffraction plane pictures," Optik (Stuttgart) **35**, 237–250 (1972).
12. N. C. Gallagher and B. Liu, "Method for computing kinoforms that reduces image reconstruction error," Appl. Opt. **12**, 2328–2335 (1973).
13. F. B. McCormick, "Generation of large spot arrays from a single laser beam by multiple imaging with binary phase gratings," Opt. Eng. **28**, 299–304 (1989).
14. M. P. Dames, R. J. Dowling, P. McKee, and D. Wood, "Efficient optical elements to generate intensity weighted spot arrays: design and fabrication," Appl. Opt. **30**, 2685–2691 (1991).
15. E. G. Johnson and M. A. Abushagur, "Microgenetic-algorithm optimization methods applied to dielectric gratings," J. Opt. Soc. Am. A **12**, 1152–1160 (1995).
16. J. P. Kirk and A. L. Jones, "Phase-only complex-valued spatial filter," J. Opt. Soc. Am. **61**, 1023–1028 (1971).
17. T. J. Suleski and D. C. O'Shea, "Gray-scale masks for diffractive-optics fabrication: I. Commercial slide imagers," Appl. Opt. **34**, 7507–7517 (1995).
18. D. C. O'Shea and W. S. Rockward, "Gray-scale masks for diffractive-optics fabrication. II. Spatially filtered halftone screens," Appl. Opt. **34**, 7518–7526 (1995).
19. B. Wagner, H. J. Quenzer, W. Henke, W. Hoppe, and W. Pilz, "Microfabrication of complex surface topographies using grey-tone lithography," Sens. Actuators A **46–47**, 89–94 (1995).
20. T. R. Jay and M. B. Stern, "Preshaping photoresist for refractive microlens fabrication," Opt. Eng. **33**, 3552–3555 (1994).
21. *Shipley Corporation Microposit Products Catalog*, Marlboro, Mass.
22. I. S. Gradshteyn and I. M. Ryzhik, *Table of Integrals, Series, and Products* (Academic, New York, 1980), p. 318, Eq. (3.382.4).